System Comparison Procedures for
Automatic Target Recognition Systems

THESIS

Anne E. Catlin
Second Lieutenant, USAF

AFIT/GOR/ENS/97M-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GOR/ENS/97M-03

System Comparison Procedures for Automatic Target Recognition Systems

THESIS

Anne E. Catlin
Second Lieutenant, USAF

AFIT/GOR/ENS/97M-03

AFIT/GOR/ENS/97M-03

System Comparison Procedures for Automatic Target Recognition Systems

THESIS

Presented to the Faculty of the Graduate School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Operations Research

Anne Elizabeth Catlin, B.S.
Second Lieutenant, USAF

March 1997

Disclaimer: The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

THESIS APPROVAL

Student: Anne E. Catlin, Second Lieutenant, USAF　　Class: GOR-97M

Title: System Comparison Procedures for Automatic Target Recognition Systems

Defense Date: 21 February 1997

Committee:	Name/Title/Department	Signature
Advisor	Kenneth W. Bauer, Jr. Professor Department of Operational Sciences Air Force Institute of Technology	Kenneth W Bauer Jr.
Reader	Edward F. Mykytka Acting Department Head Department of Operational Sciences Air Force Institute of Technology	Edward F. Mykytka

Acknowledgments

First and foremost, I must thank my advisor, Dr. Ken Bauer, for his patience, guidance, patience, encouragement, and patience throughout the thesis process. I particularly appreciate him letting me run with the research, but not letting me run off any cliffs (without a bungee cord, anyway).

My reader, Dr. Ed Mykytka, provided invaluable (though often frustrating) suggestions and perspective with the manuscript editing. I particularly appreciate his patience when I chose to torture him with hopeless new ideas on Friday afternoons.

Thanks also go to Dr. Doug Montgomery for reading the thesis and teaching me some design of experiments over the phone. His comments provided a bigger-picture perspective and let me know that I was actually doing something useful.

Several members of the MSTAR team at Wright Laboratory provided essential insight into the MSTAR system and SAR data. Thanks are due to Maj Tom Burns, the MSTAR Project Manager, for bringing the topic to my attention and sending me places to learn about MSTAR. I also appreciate the support of Dr. Tim Ross, Dr. Ron Dilsavor, Dr. John Mossing, and the rest of the MSTAR team SEQAL denizens, who helped me run MSTAR, obtain data, learn about ATR, and otherwise catch enough clue to write this thesis.

Finally, I must thank my parents for their support and encouragement; Ken Haertling for sitting through the first defense dry-run and keeping me sane throughout the thesis process; and my roommates Christine and Kim for cooking and doing laundry when I was stuck in my study writing for days on end. They were great family to have during the few hours when I wasn't chained to my computer.

Anne E. Catlin

Table of Contents

List of Figures

Figure *Page*

List of Tables

Table *Page*

AFIT/GOR/ENS/97M-03

Abstract

Estimating the performance of an automatic target recognition (ATR) system in terms of probability of successful target identification involves extensive image collection and processing, which can be very time-consuming and expensive. Therefore, we investigate the Wald sequential test for the difference in two proportions as a sample size-reducing alternative to the ranking and selection procedure and confidence intervals. An analysis of the test parameters leads to a practical methodology for implementing the Wald test for fairly comparing two systems, based on experimental goals. The test is also modified with the multiple sequentially rejective Bonferroni procedure for the multiple pairwise comparison of more than two systems, and two sampling schemes for different experimental goals are discussed.

The test methodology was applied to actual data to compare different configurations of the Moving and Stationary Target Acquisition and Recognition (MSTAR) System with good results. In a two-system comparison with real data, the Wald test required an average of about one sixth as many samples as a confidence interval to choose the superior system, and about one fifth as many samples as ranking and selection. To compare four systems with simulated data, the Wald test usually needed only one third as many samples as multiple pairwise confidence intervals to detect specified differences between the proportions, and one half as many samples as required by ranking and selection. These sample size savings demonstrate that the Wald sequential procedure with the modifications described in this thesis is a useful alternative to comparing proportions with confidence intervals, particularly when data is expensive.

System Comparison Procedures for Automatic Target Recognition Systems

I. Introduction

1.1 The Automatic Target Recognition System Performance Estimation Problem

Over the last fifty years, scientists and engineers have developed increasingly complex electronic machines designed to think like humans, but more quickly and efficiently. Long gone are the days of room-sized computers performing only simple arithmetic; today, computers use complex logic to make decisions, and can also interpret spoken language and "see" visual images. Advanced software can enable a computer to identify features in an image from a variety of sources, and actually recognize objects from trees to trucks. Software exists for the processing of infrared and laser radar imagery, and the automatic target recognition (ATR) community is currently working to create an efficient system for analyzing synthetic aperture radar (SAR) returns.

Engineers have developed extensive theory and algorithms for ATR, but have focused little on the testing and evaluation of their systems. Many ATR system tests have involved collecting and analyzing huge amounts of data and estimating performance parameters, such as probability of detection, without accounting for observed variability in estimates of these parameters. Further, test image sets, sample sizes, and performance estimation methods vary between developers, so that comparisons of ATR systems based on performance statistics are at best crude, and often invalid. The ATR community could thus benefit greatly from a standardized test procedure which integrates statistical theory with a common set of test images. This standardized plan concept is currently known as the “honest broker” approach [2].

Prospective users of the “honest broker” plan include ATR developers worldwide using imagery from many different sensor types. For example, a team assembled by the Wright Laboratory Avionics Directorate (WL/AA) at Wright-Patterson Air Force Base, Ohio, is currently developing a SAR ATR system named MSTAR (Moving and Stationary Target Acquisition and Recognition). WL needs to estimate MSTAR performance in terms of probability of target identification, and determine with 95% confidence whether the system meets specifications. MSTAR is designed to identify over 1.6×10^{11} different target scenarios, and any valid estimate must represent performance across the entire scenario space. Testing the system on all the possible scenarios is impractical; WL has neither the time nor the data to perform such a vast test, and data collection is expensive and time-consuming. Since SAR imagery requires considerable processing time, and WL needs to estimate performance for eight possible system configurations and choose the best within a limited time frame, we want to minimize the number of necessary test images. Chapter 2 explains these analysis goals in depth. The current WL Performance Evaluation Team technique is to "randomly" select about 300 representative images from the scenario space, and calculate confidence intervals to estimate performance.

1.2 Objective

We applied the Wald sequential test for the difference of two proportions to the fair comparison of ATR systems, as a sample size-reducing alternative to confidence intervals. A methodology for implementing the test is explained. A modified Bonferroni approach is then developed to compare more than two systems while maintaining the overall desired level of significance for the tests, and the methodology for the multi-system test is outlined. The test methodologies are then applied to compare different MSTAR configurations, to choose the optimal system configuration.

1.3 Scope and Summary

After an introduction to MSTAR in Chapter 2, we discuss the challenges presented by binomial data and introduces the Wald test sequential test for the difference of two proportions in detail in Chapter 3. Chapter 4 begins with analyses of all Wald test parameters, and includes derivations of the appropriate parameter relationships which force the comparisons to be fair. A step-by-step guide for parameter selection and test implementation, in accordance with experimental goals, for the fair comparison of two system with the Wald test follows. The modified sequentially rejective Bonferroni procedure, explained in Chapter 3, is then incorporated into the Wald test to create a valid methodology for the simultaneous fair comparison of four systems. The chapter concludes with a discussion of sampling procedures for data input to the test.

Chapter 5 applies the methodology of Chapter 4 to MSTAR data to meet WL test goals. Wright Laboratory outlined the following performance evaluation goals (the terms "peaks" and "regions" will be defined in Chapter 2):

1. Choose the best of four system configurations which analyze image peaks.
2. Choose the best of four system configurations which analyze image regions.
3. Determine whether MSTAR works better when analyzing the peaks or regions version of a particular configuration.

We address all three of goals in Chapter 5, beginning with the third goal, which involved a simpler 2-system comparison rather than the more complex 4-system comparison of the first two goals. Given test parameters set to WL specifications, the Wald test chose the regions configuration as superior to the peaks configuration in nearly 100% of Wald test runs with the data randomly reordered each time, using about a quarter of the samples necessary to compute confidence intervals of similar accuracy. The data sets provided for the other two testing goals were too small for conclusive testing, so simulated data was used instead, generated from estimated probabilities of identification from the WL data sets. The Wald test for 4-system comparisons developed in Chapter 4 successfully chose the best systems using about one third of the samples necessary to compute confidence intervals.

II. The MSTAR System

This chapter provides an introduction to the Moving and Stationary Target Acquisition and Recognition (MSTAR) System, an example of an ATR system which presents an ideal application for sequential testing. We include a primer on synthetic aperture radar, and discuss the factors which affect the performance of an ATR system.

2.1 MSTAR System Design and Configuration Competition

The Moving and Stationary Target Acquisition and Recognition (MSTAR) System is a model-based approach to automatic target recognition of SAR imagery. Previous solutions to the SAR ATR problem relied on impractically vast data libraries of targets at numerous angles and configurations. The model-based algorithm relies on computer-generated templates for matching an identity to each image, based on only a small data library of stored actual SAR imagery of targets.

The system consists of six modules [11]:

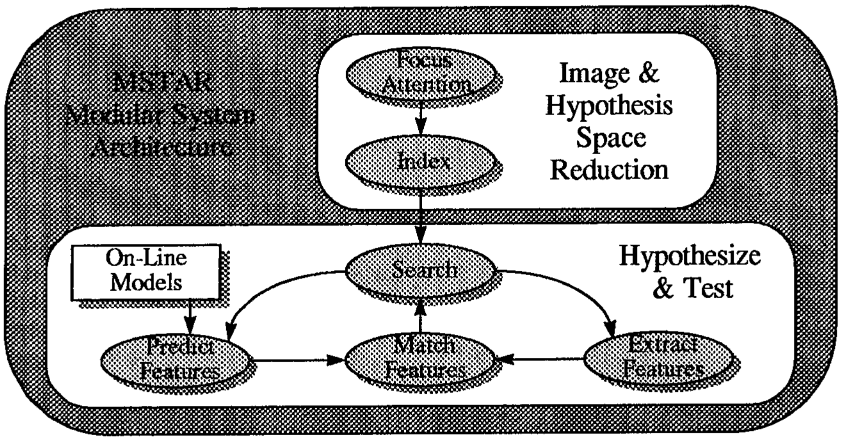

Figure 2.1 MSTAR system architecture [11]

- *Focus of Attention (FOA):* From SAR image input, identifies regions of interest in the image (ROIs) which may contain targets
- *Indexing (IX):* For each ROI, generates a list of potential "hypotheses" from the stored imagery database, which serve as initial guesses of the target's identity
- *Search (S):* Investigates hypotheses and searches for model improvements; acts as a central node in the PEMS loop (Predict, Extract, Match, Search; see below)
- *Predict (PR):* Based on the hypothesized model, predicts features which may occur in the target image
- *Feature Extract (FE):* Identifies and extracts peaks and regions features from the ROI (explained in Section 2.4)
- *Match (M):* Matches extracted features with the predicted target features

The last four modules form an iterative loop, called PEMS (predict, extract, match, search). This loop explores the hypothesis space for the hypothesis which matches the

ROI with the highest possible level of confidence. The Search module outputs a target report listing the specific class and identity of the target (for example, a T72 tank) for each ROI.

In June 1995, DARPA (Defense Advanced Research Projects Agency) funded and WL awarded contracts for developing and integrating the system modules described above. Several of the modules were dually awarded, yielding several possible system configurations. A configuration is a specific combination of the six modules in Figure 2.1. Further, systems which extract and match image peaks can be considered different configurations than those which consider regions. Explanations of peaks and regions follow in Section 2.4.

Of the configurations which meet specifications, the best configuration will correctly identify targets in a SAR image more frequently than any other configuration. Module interaction may affect configuration performance; for example, the match module in the best configuration may not be the best stand-alone match module. WL must select the single contractors for the dually-awarded modules which are included in the best full-system configuration.

2.2 *The MSTAR Performance Estimation Problem*

In addition to selecting the best configuration, WL needs to estimate four performance parameters to verify that at least one configuration of the MSTAR system meets the following specifications for each parameter with 95% confidence:

- Probability of detection (p_D) ≥ 0.9
- Probability of correct classification (p_{CC}) given detection ≥ 0.9
- Probability of correct identification (p_{ID}) given detection ≥ 0.7
- False alarm rate $\leq 0.001/\text{km}^2$

"Correct detection" is defined as correctly declaring that a target in an ROI is, in fact, a target. A "false alarm," or incorrect detection, occurs when the system declares clutter, such as trees, as a target. "Correct classification" is defined as the system properly classifying a detected target as a member of one of five classes listed in the table below. For example, if we input an armored personnel carrier (APC) image and MSTAR reports the same, though it may not correctly identify the specific APC, we have a correct classification. "Identification," a subset of classification, is naming the specific alphanumeric target designator. A sample correct identification would occur if we enter a T72 image into MSTAR, and MSTAR reports identification of a T72.

WL wants to test the null hypothesis that $p_{ID} < 0.7$ against the alternative $p_{ID} \geq 0.7$, where p_{ID} is the probability of correct target identification. We could simultaneously test H_0: $p_{CC} < 0.9$ against H_1: $p_{CC} \geq 0.9$, where p_{CC} is the probability of correct classification. The estimates of both parameters, $\hat{p}_{ID}$ and $\hat{p}_{CC}$, are conditioned on detection and can be calculated from test data as follows [6]:

$$\hat{p}_D = \frac{\text{number of target images declared as targets}}{\text{number of target images tested}} \qquad (2.1)$$

$$\hat{p}_{CC} = \frac{\text{number of correctly classified target images}}{\text{number of detected target images}} \tag{2.2}$$

$$\hat{p}_{ID} = \frac{\text{number of correctly identified target images}}{\text{number of detected target images}} \tag{2.3}$$

This study focuses on only on p_{ID} such that all result-related experimental decisions depend on p_{ID} estimation above estimating any other performance parameter. WL wants to know whether at least one MSTAR configuration meets these specifications, but assuming that at least one system has a p_{ID} of at least 0.7, they are most interested in finding the best system. Since we focus on p_{ID}, we will assume that all targets have been detected, and we want to study system performance in identifying detected targets.

2.3 MSTAR Extended Operating Conditions (EOC's)

The MSTAR extended operating conditions (EOC's) form a set of "highly challenging, militarily realistic scenarios" which MSTAR is being designed to operate under [3]. The EOC's incorporate eight factors, each with several variations (the number of variations for each factor is included in parentheses):

1. target type (20)
2. aspect angle from sensor to target (continuous, 360°)
3. depression angle from sensor to target (continuous, all angles from 10° to 45°)
4. articulation (examples: hatches open or closed; SCUD up or down) (36)
5. target configuration (variations on target type, e.g. old vs. new T72) (16)
6. obscuration of target by barriers, trees and other objects (20)
7. layover camouflage (20)

8. netting camouflage (5).

Of the twenty MSTAR target types, the thirteen target types used in the first phase of the program are listed in Table 2.1.

Table 2.1

MSTAR EOC target types [6]

Main Battle Tank (MBT)	Armored Personnel Carrier (APC)	Self-Propelled Gun (SPG)	Truck (T)	Mobile Missile Launcher (MML)
T72† M1	BMP2† M2† *M113* _BTR60*_ BTR70*	M109 M110	M548 M35* _HMMWV*_	SCUD*

* Wheeled vehicles. All others have tracks.

† Three examples available in the collected MSTAR data. Only one example available of all others.

The version of MSTAR tested in this study was *not* developed to identify the three italicized targets above. They act as "confusers," or targets which the system should declare in the "other" category.

In addition to the target type, engineers have identified the aspect and depression angles of the target from the sensor as the two other most important factors in the critical factor set. These angles represent orientation of the SAR sensor with respect to the target, as seen in Figure 2.2, where ϕ_A is the aspect angle and ϕ_D is the depression angle.

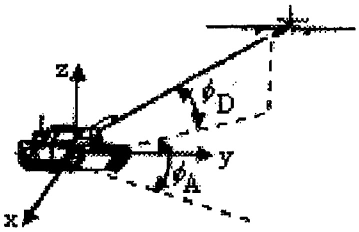

Figure 2.2. Aspect and depression angles [3].

WL randomly drew a representative set of images from the space of all target types, aspect angles (sometimes called azimuth), and depression angles for use in estimating MSTAR performance. Each target (T) at each aspect (A) and depression (D) angle has a different p_{ID}, since SAR imagery characteristics, including broadside flash, can make a target more or less difficult to detect at different angles. These characteristics are explained in Section 2.4. Henceforth, a T, A, D scenario refers to a specific target at a certain aspect angle and a certain depression angle, such as a T72 at 45° aspect and 30° depression. Therefore, we define p_{ID} for a scenario as the probability of identification for a given T, A, and D. All of the other factors and exterior conditions are lumped into a variable which we will call environment (henceforth E), and which accounts for the randomness in the estimate of p_{ID} for each scenario, $\hat{p}_{ID}$. We further define the overall p_{ID} for an ATR system as the mean of all scenario p_{ID}'s across the full discrete range of targets, and the continuous ranges of all aspect angles and depression angles from 15° to 45°. We will use p_{ID} as the single performance measure for comparing the MSTAR configurations.

Note that the p_{ID} of a T, A, D scenario is not binary and may lie anywhere in the interval [0,1], since MSTAR may correctly identify one image of a scenario, but may not correctly identify a different example of the same scenario. Most military targets, particularly tanks, have dents or homemade additions which will alter the SAR image; in one example, a crew had bolted a hibachi to the back of their tank, which had square metal corners and thus changed the appearance of the SAR return. If we process one specific ROI through MSTAR multiple times, the system will always produce the same target identification report, but a different ROI of the same T, A, D scenario may be differently identified by MSTAR.

2.4 Synthetic Aperture Radar (SAR)

Though ATR systems have been built using a variety of sensors, MSTAR uses SAR imagery. SAR sensors collect data while moving around an object, while the object is moving, or both; therefore, image quality can vary with a sensor's orientation to a target in terms of aspect and depression angle [10:14]. Synthetic aperture radar simulates a large antenna by scanning an object with a fanlike beam from a small antenna, and processing the resulting radar return, which consists of phase-shift data. Sensor range does not affect resolution, so the sensor vehicle can operate at considerable range from the object [13]. In military applications, this means that an aircraft with a SAR sensor can collect imagery while maintaining a safe distance from a target. SAR sensors usually use microwaves, which allows all-weather surveillance of potential targets [13]. SAR images look as if the sensor has shone light on a contoured, mirrored surface. Imagine walking into a dark room full of mirrored objects with a miner's light on your forehead; what you would see resembles a SAR image, with large and small bright spots and dim regions. Sharp corners, such as the inside corner of a truck bed, will produce large bright spots when imaged from certain angles, while the flat area on the front of a tank may appear as a dim or dark region between bright spots. Images consist of a variety of light and dark spots in arrangements corresponding to the imaged target. The bright spots, known as "peaks," and the regions comprise the two types of features in SAR imagery. The MSTAR system can identify targets based on either the peaks or regions, thus adding feature type to the factor set. Figure 2.3 is a sample SAR image of an M-60 in which the peaks are clearly visible.

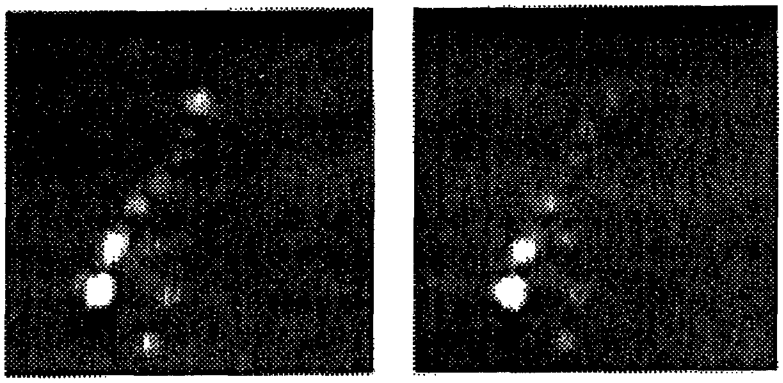

Figure 2.3. Synthetic SAR imagery of M-60 at 20° depression and -123° (left image) and -120° (right image) aspect [3]

At some aspect angles, particularly 0°, 90°, 180°, 270°, the broadside flash effect can occur. This effect results when the many bright peaks visible from a cardinal angle overlap and blur into a flash. This effect makes image analysis very difficult for a trained human, let alone an ATR system.

The mathematics of SAR are beyond the scope of this thesis, but the resolution of the sensor data affects the levels of aspect which we may test. The SAR sensor used to collect MSTAR data has resolution of 1 foot and a frequency of 10 GHz, so that sensor data collected 3° of aspect apart can be considered independent. The following angle resolution calculation below demonstrates this, where λ is wavelength, f is frequency, ϕ_A is aspect angle, and c is the speed of light (recall from physics that $\lambda f = c$):

$$resolution = \frac{\lambda}{2\sin\phi_A} \tag{2.4}$$

$$\phi_A \approx \sin\phi_A = \frac{c}{2f\,(resolution)} = \frac{2.9979\times10^8}{2(10\times10^9)(0.3048)}$$

$$\phi_A = 0.0492 = 2.818^{\circ} \approx 3^{\circ}.$$

Therefore, we can collect independent performance data at every three degrees of aspect around the target, giving us a maximum of $(360^{\circ}/3^{\circ}) = 120$ possible aspect angle levels [5] for each pass around the target at each depression angle. If the sensor platform circles the target many times, independent imagery can be collected at all aspect angles. The MSTAR EOC list also includes all depression angles between 10° and 45°; however, only two are available for this experiment, restricting our choice of levels. The experiments in Chapter 5 were thus performed on data which is not truly representative of the whole depression angle space, but we will treat them as such for demonstrative purposes.

III. Binary Data and Sequential Testing

This chapter describes some simple statistics and techniques for analyzing binary data, and introduces the Wald test for the difference in two proportions. We consider the Wald test since it generally requires fewer sample images to compare two ATR systems than are required by conventional statistics, and is thus more efficient in most cases.

3.1 Binary Responses

To estimate p_{ID}, for example, we must process enough images through a given MSTAR configuration to estimate performance with sufficient accuracy, per WL judgment. The outcome of each run consists of a correct or incorrect identification; in other words, the experiment yields a *binary response*.

Estimating the probability of success p of a binomial distribution involves averaging n observations x_i, where $x_i \in \{0,1\}$ (0 for incorrect identification and 1 for correct identification):

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

The variance $\frac{\hat{p}(1-\hat{p})}{n}$ of the estimate $\hat{p}$ is dependent on both $\hat{p}$ and the number of independent samples n. Therefore, confidence interval widths vary with both parameters

[12:417], as seen in Equation (3.2), which assumes large sample sizes and thus uses $z_{1-\alpha/2}$ instead of $t_{1-\alpha/2}$ as the critical statistic.

$$\text{for large } n\text{: } \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3.2}$$

We generally need very large sample sizes to obtain reasonably narrow interval widths, but also want to reduce data requirements as much as possible. To compare two systems, we can either compare the interval calculated via Equation (3.2) for two systems, or we can use an interval for the difference $|p_1 - p_2|$, henceforth called a difference interval. The difference interval allows us to choose a superior system with fewer samples than comparing two intervals, but still needs large sample sizes when the difference is small. As in the construction of Equation (3.2), we require independent sampling for estimation of $\hat{p}_1$ and $\hat{p}_2$. We calculate difference intervals for large sample sizes via

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \tag{3.3}$$

When constructing intervals, we want the data to represent the spectrum of aspect angles and depression angles for all of the targets, since p_{ID} may differ for different targets at different angles. The binomial confidence interval method assumes constant variance for all observations, and since $p_{ID}(1 - p_{ID})$ is different for different scenarios, the MSTAR case violates this assumption. Unless we sample data to estimate p_{ID} for each configuration independently and thus treat the data as samples from a binomial population with parameter p_{ID}, we must derive and use a confidence interval which accounts for variation in the scenario p_{ID}‘s. Given binary data from a set of scenarios which represents performance across the entire aspect, depression, and target space, we can construct the confidence interval in Equation (3.4), derived in Appendix A, where *i, j,* and *k* represent target, aspect, and depression, $\hat{P}^{ijk}$ is the estimated probability of identification for a scenario with target *i*, aspect *j*, and depression *k*, and N_{ijk} is the number of independent observations of scenario *ijk*. P_{ijk} is the probability of acquiring scenario *ijk*; for example, guerrillas in underdeveloped countries may have more trucks than Scuds, so the probability of acquiring a Scud would be lower than that for a truck. The MSTAR design initially assumes that all targets are equally likely, to be mission-nonspecific.

$$\sum_i \sum_j \sum_k \hat{P}^{ijk} P_{ijk} \pm z_{1-\alpha/2} \sqrt{\sum_i \sum_j \sum_k \hat{P}^{ijk}\left(1-\hat{P}^{ijk}\right)\frac{P_{ijk}^2}{N_{ijk}}} \qquad (3.4)$$

This confidence can be used if sufficient data for a representative scenario set is available. Unlike Equations (3.2) and (3.3), this interval accounts for the variation in observed system $\hat{p}_{ID}$ caused by T, A, D and E. A difference interval can also be constructed from this formulation. The choice of confidence interval depends on the sampling method, which is discussed in Section 4.2.

3.2 Ranking and Selection Techniques

When we want compare probabilities of success for different systems but prefer a procedure which uses fewer samples than confidence intervals, we can use ranking and selection procedures. The goal of this technique is to "select the population with the largest p value" from a set of k binomial populations with ordered p-values such that $p_{[1]} \leq p_{[2]} \leq \ldots \leq p_{[k]}$ [7:105]. Note that the procedure does not aim to specifically estimate the probabilities, but merely to identify which population has the highest p-value. The procedure is based on the distance measure δ, where $\delta = p_{[k]} - p_{[k-1]}$, the difference between the highest and second-highest p-values. The analyst must choose a value δ^*, which represents the indifference region for $p_{[k]} - p_{[k-1]}$, and a confidence level P^*, which is the probability of correctly selecting the highest p. To relate this to classical hypothesis testing, we may consider P^* as analogous to $1 - \alpha$. Tables included in the appendices of Gibbons et al. provide the required common sample size n to correctly select the best system with a probability of "at least P^* for all $\delta \geq \delta^*$" [7:107].

The test is performed by collecting n samples from each population and calculating the proportion of successes, then ordering those proportions. Mathematically, we estimate proportion p_j for population j as $\hat{p}_j = x_j/n_j$, where x_j is the number of successes in the n_j total samples collected from population j. Provided that n independent samples have been collected from each population, we may conclude with a $(1 - P^*)\%$ chance of error that we have correctly identified the highest proportion.

To illustrate this procedure, suppose we want to find the highest p from a set of four populations with 90% confidence, such that the difference between the highest and second highest p is 0.10. Therefore, $k = 4$ and $(\delta^*, P^*) = (0.10, 0.90)$. We would refer to Table E.1 of Gibbons et al. to find the required common sample size, an excerpt of which appears in Table 3.1.

Table 3.1

Smallest integer sample size n needed to satisfy the (δ^*, P^*) requirement in selecting the binomial population with the largest probability [7:425]

$k = 4$

	P^*			
δ^*	.80	.85	.90	.95
.05	359	458	601	850
.10	90	114	150	212
.15	40	51	67	94
.20	23	29	38	53

The underlined entry in the table shows that we need 150 samples of each population to calculate proportions which we can order to find the best population with 95% confidence.

The test allows for ties for first place such that when the two highest p-values are equal, one of the populations is selected at random. The random selection preserves the maximum probability of error P^* so that no correction to the confidence level is necessary [7:108].

If the desired δ^* or P^* values for an experiment do not appear in Table E.1 of Gibbons et al., we may use the figures of Appendix F of the same reference to calculate the required sample size. If δ^* or P^* is beyond the range of those figures, we may try to extrapolate from the figures [7:427], or may use an approximation if we want to detect small differences in the probabilities with a high level of confidence.

In cases where we want a small δ^* and a large P^*, we may use a normal approximation to this binomial problem. This tests uses different tables and a different sample size calculation. In this case, we approximate $n = (1 - \delta^{*2})(\tau_t/2\delta^*)^2$, where τ_t is found in Table A.1 of Gibbons et al. based on P^* and the number of systems being compared k. An excerpt of this table appears in Table 3.2.

Table 3.2

Values of τ_t for fixed P^* [7:400]

	P^*		
k	.900	.950	.975
2	1.8124	2.3262	2.7718
3	2.2302	2.7101	3.1284
4	2.4516	2.9162	3.2220
5	2.5997	3.0552	3.4532

For example, in the MSTAR case, we will compare 2 and 4 systems with $\alpha = 0.05$ and $\delta^* = 0.03$. To perform this selection test, we need to collect $n = (1 - \delta^{*2})(\tau_t/2\delta^*)^2 = (1 - 0.03^2)(2.3262/0.06)^2 = 1502$ samples from each system for the 2-system comparison, and $(1 - 0.03^2)(2.9162/0.06)^2 = 2361$ samples for a 4-system comparison. The difference intervals described in Section 3.1 require about 1800 and 3300 samples respectively to perform the same test, as shown later, so the ranking and selection procedure provides some savings over the confidence interval comparison technique in the MSTAR case.

Sample sizes for the ranking and selection procedure for choosing the best of 2, 3, or 4 systems with 95% confidence appear in Table 3.3.

Table 3.3

Minimum sample sizes for the ranking and selection procedure with $P^* = 0.95$

	Number of Systems Compared		
δ^*	2	3	4
0.01	13527	18360	21259
0.02	3381	4589	5314
0.03	1502	2039	2361
0.04	845	1146	1327
0.05	540	733	849
0.06	375	509	589
0.07	275	373	432
0.08	211	286	331
0.09	166	225	261
0.10	134	182	211
0.11	111	150	174
0.12	93	126	146

Though the ranking and selection sample sizes improve upon the confidence interval sample sizes, sequential tests may require even fewer samples. Since image processing can be time-consuming and image collection is very expensive, we prefer to compare ATR systems with the smallest possible number of samples needed to choose one system as statistically significantly better than another. We will therefore explore sequential testing, keeping in mind that we can terminate a sequential test and use another technique on the data collected if the sequential test fails to terminate within the sample sizes required by the fixed-sample-size techniques.

To address the MSTAR performance specification of $p_{ID} \geq 0.7$, we assume that at least one MSTAR configuration will meet the specification, and can verify this with a single confidence interval after comparing the systems.

3.3 Sequential Testing

Abraham Wald developed several sequential tests for making acceptance or rejection decisions of hypotheses with the minimum number of trials [19]. Some of these tests specifically concern binary data. For example, the sequential probability ratio test reduces the expected number of samples required to determine whether a proportion p is equal to one of two hypothesized proportions, p_0 or p_1, as compared to the sample sizes required by confidence intervals or ranking and selection. Appendix B contains a description of this test. Wald also derived a procedure to test for a difference in the probabilities of success for two processes, p_1 and p_2, such that one produces successful trials (i.e., a value of 1) significantly more often than the other. While we could simply test two systems on a large enough set of representative scenario data to be able to calculate confidence intervals as in Equation (3.3) which detect a difference, or perform the ranking and selection procedure, the Wald test cuts the runs required for system comparison by a large amount, particularly when the difference $|p_1 - p_2|$ is less than 0.05. Wald noted "that these sequential tests usually lead to average savings of about 50 per cent in the number of trials as compared with" confidence intervals [19:36], which can translate into significant cost savings in some experiments. Also, recall from Section 2.3 that the ROI p_{ID}'s are not constant, so the variances $p_1^i(1 - p_1^i)$ and $p_2^i(1 - p_2^i)$ are also not constant, and thus binomial confidence intervals as in Equation (3.2) are not truly valid when we cannot independently sample from p_1 and p_2. In the case of the unequal p_1^i and p_2^i across the range of i, the Wald test is a "correct procedure," while binomial confidence intervals are theoretically inapplicable [19:109].

A description of the Wald $|p_1 - p_2|$ test follows in order of implementation, starting with necessary parameter selection, and proceeding to actual execution.

3.3.1 The Wald sequential test for $|p_1 - p_2|$. In many industrial examples, a given process is installed and operating, and the company must determine whether to replace it with a new process, usually at some cost. The hypothesis for Wald's test is as follows.

$$H_0: p_1 \geq p_2$$

$$H_1: p_1 < p_2$$

The null hypothesis supposes that existing process 1 is as least as good or better than new process 2, while the alternate hypothesis states that process 2 is better, assuming that we prefer high values of p.

The test is based on the ratio of the "efficiencies" of the two processes. Wald defines "efficiency" k as the ratio of successes (ones) to failures (zeros) such that for any process, $k = p/(1-p)$, where p is the true probability of success. The value u is the ratio of efficiencies for two processes:

$$u = \frac{k_2}{k_1} = \frac{p_2(1-p_1)}{p_1(1-p_2)} \qquad (3.5)$$

Inspection shows that if $u < 1$, process 1 is better; further, if $u = 1$, the processes are equally efficient, and if $u > 1$, process 2 is better.

To implement the test, we must first set parameters u_0 and u_1. Wald explains the procedure for choosing these values in terms of manufacturing processes:

> "...select two values of u, u_0 and u_1 say, such that the rejection of process 1 in favor of process 2 is considered an error of practical importance whenever the true value of $u \leq u_0$, and the maintenance of process 1 is considered an error of practical importance whenever $u \geq u_1$. If u lies between u_0 and u_1, the manufacturer does not particularly care which decision is made." [19:110]

Since u-space is less than intuitive, we may choose u values more easily by investigating u in terms of p_1 and p_2. Table 3.4 displays of u values for various p_1 and p_2 values.

Table 3.4

Values of u for the full range of possible p_1 and p_2

p_1/p_2	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
0.10	0.00	1.00	2.25	3.86	6.00	9.00	13.50	21.00	36.00	81.00
0.20	0.00	0.44	1.00	1.71	2.67	4.00	6.00	9.33	16.00	36.00
0.30	0.00	0.26	0.58	1.00	1.56	2.33	3.50	5.44	9.33	21.00
0.40	0.00	0.17	0.38	0.64	1.00	1.50	2.25	3.50	6.00	13.50
0.50	0.00	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00
0.60	0.00	0.07	0.17	0.29	0.44	0.67	1.00	1.56	2.67	6.00
0.70	0.00	0.05	0.11	0.18	0.29	0.43	0.64	1.00	1.71	3.86
0.80	0.00	0.03	0.06	0.11	0.17	0.25	0.38	0.58	1.00	2.25
0.90	0.00	0.01	0.03	0.05	0.07	0.11	0.17	0.26	0.44	1.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

As seen in the table, u can assume values from zero to very large; values in the $p_1 = 0.0$ row and $p_2 = 1.0$ column are infinite and were therefore omitted.

Since u ranges significantly with p_1 and p_2, we may find a table with higher resolution in a specific area of interest more useful. For this, we need an educated guess or reference point of the true values of p. For example, in the MSTAR case, $p_{ID} > 0.7$ according to system specifications. Therefore, we are most interested in detecting differences in p_1 and p_2 when both are near 0.7. Table 3.5 contains u-values in this neighborhood.

Table 3.5

Values of u for selected p_1 and p_2 from 0.66 to 0.75

p_1/p_2	**0.66**	**0.67**	**0.68**	**0.69**	**0.70**	**0.71**	**0.72**	**0.73**	**0.74**	**0.75**
0.66	1.00	1.05	1.09	1.15	1.20	1.26	1.32	1.39	1.47	1.55
0.67	0.96	1.00	1.05	1.10	1.15	1.21	1.27	1.33	1.40	1.48
0.68	0.91	0.96	1.00	1.05	1.10	1.15	1.21	1.27	1.34	1.41
0.69	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.21	1.28	1.35
0.70	0.83	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.22	1.29
0.71	0.79	0.83	0.87	0.91	0.95	1.00	1.05	1.10	1.16	1.23
0.72	0.75	0.79	0.83	0.87	0.91	0.95	1.00	1.05	1.11	1.17
0.73	0.72	0.75	0.79	0.82	0.86	0.91	0.95	1.00	1.05	1.11
0.74	0.68	0.71	0.75	0.78	0.82	0.86	0.90	0.95	1.00	1.05
0.75	0.65	0.68	0.71	0.74	0.78	0.82	0.86	0.90	0.95	1.00

Suppose that we are interested in detecting a difference of 0.03 between p_1 and p_2. We can look along the underlined diagonals in Table 3.5 where $|p_1 - p_2| = 0.03$. For $p_1 > p_2$, the u values are 0.86 or 0.87, and for $p_1 < p_2$, u ranges from 1.15 to 1.17. Note that $u_0 = 1/u_1$, since we want to compare the systems fairly. In a manufacturing case where process 1 with probability of success p_1 is already installed and running, we may prefer to maintain process 1 if $p_1 - p_2 > 0.03$, but we will only install system 2 if $p_2 - p_1 > 0.05$. If $p_2 - p_1 = 0.04$, we will maintain process 1. In such a situation, the u-values will not necessarily be reciprocals, but in the MSTAR case, we want to compare the systems fairly. Further argument for setting $u_0 = 1/u_1$ appears in Section 4.3.

After setting u_0 and u_1, we also choose the values α and β to reflect risk tolerance. In many statistical tests, α and β are the desired significance level and 1 minus the power of the test respectively. The significance level α is the probability of a type I error, the case in which the test rejects the null hypothesis when the null is actually true. β is the probability that the test accepts the null when the alternate hypothesis is correct, also known as Type II error [8:224]. Appropriate values of α and β depend on the required accuracy of the test and the resources available, since increased accuracy usually increases the cost of the experiment. However, Wald also defines α and β as risk tolerances for this test. Parameter α is an upper bound on "the probability of rejecting process 1...whenever $u \leq u_0$," and parameter β is an upper bound on "the probability of maintaining process 1...whenever $u \geq u_1$" [19:110]. Parameters α and β represent risk tolerances for choosing the wrong process, as well as test significance and power. Also, the actual probability of incorrectly choosing process 1 could be less than α for some u-values, for example, since α is an upper bound.

The last step in setting up the test is to ensure that the data is in paired format. In other words, the two sets of zeros and ones from each process must be paired in the form (c_i, d_i), such that c_i is the i^{th} data point from process 1, and d_i is the i^{th} result from process 2. In the MSTAR case, a data point is the result of the system processing a specific ROI, and holds a value of one if the system correctly identified a target, and zero if it incorrectly identified the target. To maintain the fairest comparison between the configurations, we could require that c_i and d_i be the processing results from the exact same ROI, where each ROI has a different p_{ID} depending on T, A, D and E. This blocks on the effects of T, A, D, and E, so only system performances are compared, and is mathematically correct since the Wald test allows for tests in which p_{ID} is nonconstant between samples [19:109]. Alternatively, we can randomly draw an ROI for c_i and another for d_i, ignoring the effects of T, A, D and E and treating the data as random draws from a binomial distribution with a constant overall system p_{ID}. The choice of sampling scheme depends on the experimental goals, and will be discussed further in Chapter 4.

Starting with the first data pair and proceeding through the set, we increment a counting variable, called t_1, each time a pair (1,0) appears, and increment counting variable t_2 for each (0,1). The test ignores all (0,0) and (1,1) pairs. Further, t is the total number of (0,1) and (1,0) pairs, such that $t = t_1 + t_2$. The following equations provide critical values of t_2 for choosing the superior process [19:111]:

$$\text{lower bound: } \frac{\log \frac{\beta}{1-\alpha}}{\log u_1 - \log u_0} + t \frac{\log \frac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} \tag{3.6}$$

$$\text{upper bound: } \frac{\log \frac{1-\beta}{\alpha}}{\log u_1 - \log u_0} + t \frac{\log \frac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} \tag{3.7}$$

If t_2 falls below the lower bound for any value of t, we conclude that process 1 is better than process 2. If t_2 exceeds the upper bound, process 2 is better than process 1. If t_2 remains between the bounds, we continue testing.

We can depict this procedure in graphical form, plotting the values of the bounds as functions of t:

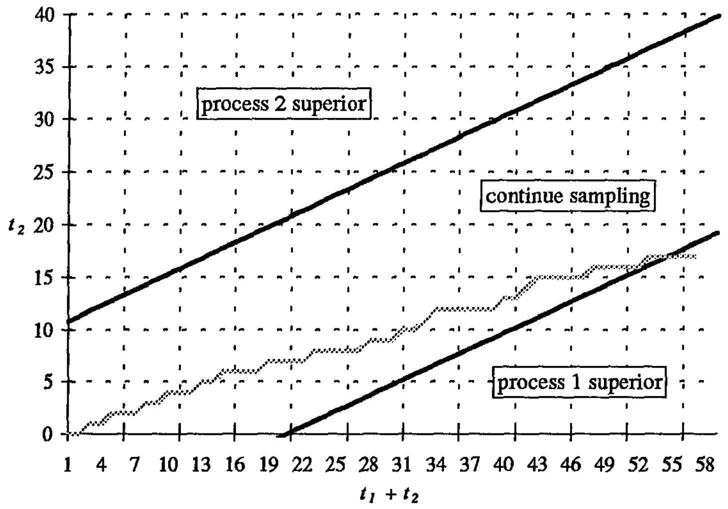

Figure 3.1. Wald Test of H_0: $p_2 \leq p_1$

When t_2 breaks through a boundary, conclusions can be drawn as labeled.

3.3.2 Comparing multiple systems. In comparing more than one process, we could perform several of these tests simultaneously with the same critical values. For example, in the MSTAR case, we want to choose the best of systems A, B, C and D. The null hypothesis is now $p_A = p_B = p_C = p_D$, where p_X is the probability of identification for system X. The alternate hypothesis is that any p_X is higher than any other p_X. We construct the $\binom{4}{2}$ = 6 simultaneous pairwise tests with the following null hypotheses:

$$\begin{gathered} p_A = p_B \\ p_A = p_C \\ p_A = p_D \\ p_B = p_C \\ p_B = p_D \\ p_C = p_D \end{gathered}$$

Further suppose that when $t_1 + t_2 = 26$, the fifth test rejects the null hypothesis and finds that $p_B > p_D$. Though we have not disproven the hypothesis that $p_A = p_D$ and $p_C = p_D$, we can discard D, as we have shown that it cannot be the best system.

When performing multiple pairwise comparisons like those described in the previous example, we must modify our choice of α. If the maximum probability of type I error of each pairwise comparison is 0.05, then the probability of correctly choosing the best configuration in all six tests could be as poor as $(1 - 0.05)^6 = 0.735$ if the tests were independent. Since probabilities of selecting the correct configuration of less than $(1 - \alpha)$ are unsatisfactory by definition, we can apply the Bonferroni inequality and use a level of significance α/M for each test, where M is the number of pairwise comparisons. The Bonferroni inequality states that the sum of the maximum probabilities of type I error for all of the pairwise tests A_i is at least as great as the joint significance for all of the pairwise comparisons together

[17:153]. Mathematically, if we perform M simultaneous tests and let A_i denote the event that test i results in a type I error, then

$$P\left(\bigcap_{i=1}^{M} \overline{A_i}\right) \geq 1 - \sum_{i=1}^{M} P(A_i). \tag{3.8}$$

Therefore, using a significance level of $P(A_i) = \alpha/M$ for each pairwise test provides a conservative approximation of the desired joint maximum probability of type I error of α

for all of the multiple comparisons. The same procedure may be applied to parameter β to achieve the desired maximum joint probability of type II error.

To increase the power of multiple pairwise comparison tests, Holm [9] proposed a sequentially rejective Bonferroni (SRB) procedure which maintains α as the overall test significance. Rather than using α/M for all tests, Holm uses $\alpha / (M - j)$, where j is the number of hypotheses which have been rejected so far, or pairwise tests which have terminated. For example, in the $\binom{4}{2}$ = 6 pairwise tests case with a desired overall maximum probability of type I error of 5%, we begin with an individual test maximum probability of type I error of $\alpha/6 = 0.0083$. After one of the tests terminates, the remaining five continue with an individual maximum probability of type I error level of $\alpha/(6 - 1) = 0.01$, and so forth until the final remaining test uses $\alpha = 0.05$. For a given set of pairwise tests, this SRB procedure allows at least as many tests to conclude as the original Bonferroni procedure, and can actually resolve some of the pairwise tests which the Bonferroni procedure may deem inconclusive. (Since the Bonferroni procedure maintains α/M as the significance even to the last remaining hypothesis, the test boundaries seen in Figure 3.1 are often too far apart for the plot of a limited data set to reach either boundary if the difference between the p_{ID}'s for the final two systems is small. Therefore, a test may fail to conclude, whereas the SRB allows the boundaries to move closer together after some hypotheses are rejected or accepted, so that the same test may reach a boundary more easily and thus conclude.)

Shaffer [18] further modified the SRB to account for overlap in test hypotheses. Suppose, for example, that in the case of processes A, B, C and D above, we found that $p_B > p_D$. Under the SRB procedure, our individual test maximum probability of type I error would increase from $\alpha/6 = 0.0083$ to $\alpha/5 = 0.01$. However, logically, we are also eliminating the tests for $p_A = p_D$ and $p_C = p_D$, since we know that system D cannot be the best. The maximum number of possible true hypotheses t_j, or test outcomes, is actually reduced from six to three. Therefore, we can continue testing with $\alpha/3 = .0167$, and the individual test maximum probability of type I error under Shaffer's Modified Sequentially Rejective Bonferroni procedure (MSRB) is α/t_j. Holm proved that the SRB maintains the overall maximum probability of type I error of $\leq \alpha$, and the MSRB follows under the same proof, summarized by Shaffer, where m is the total number of true hypotheses, n is the maximum possible number of true hypotheses, and Y_i is the probability that the test statistic of hypothesis i is greater than the statistic for the observed data:

> "The basic idea behind Holm's proof is that if m hypotheses are true, *an error must occur at or before stage $n - m + 1$*. Therefore, P(no errors) $\geq$ P($Y_i \leq \alpha/m$ for some $i \in I$) $\geq 1 - \sum_{i \in I} \alpha / m = 1 - \alpha$." [18:827]

We thus may use α/t_j in place of α in Wald tests which we intend to use for multiple pairwise comparison purposes, and modify the Wald bounds to reflect changes in the individual test significance level as the test progresses. Henceforth, we refer to the Wald $|p_1 - p_2|$ test modified by the MSRB as the Wald MSRB test.

3.3.3 Expected sample size calculations. When performing the Wald MSRB or any other test which requires sampling, experimenters usually need an estimate of the data requirements for the test. For example, when testing whether a certain drug helps to accelerate the recovery of heart attack victims, doctors need to know how roughly many patients they need for the study so they can collect data as efficiently as possible.

Wald derived expected sample size equations for all of his sequential tests [19]. For the $|p_1 - p_2|$ test, the expected sample size is calculated as follows. $E_u(t)$ is the expected value of t, where t is the total number of unmatched pairs counted during testing, (i.e. $t = t_1 + t_2$, where t_1 and t_2 are the counters for (1,0) and (0,1) pairs, respectively) [19:115]:

$$E_u(t) = \frac{L(u)\left(\log\frac{\beta}{1-\alpha}\right) + (1-L(u))\left(\log\frac{1-\beta}{\alpha}\right)}{\frac{u}{1+u}\log\frac{u_1(1+u_0)}{u_0(1+u_1)} + \frac{1}{1+u}\log\frac{1+u_1}{1+u_0}} \qquad (3.9)$$

In this equation, $L(u)$ is the probability that the test will retain process 1 when that process is as at least as good as process 2 [19:113]. $L(u)$ and u are calculated as [19:114]:

$$L(u) = \frac{\log\frac{1-\beta}{\alpha}}{\log\frac{1-\beta}{\alpha} - \log\frac{\beta}{1-\alpha}} \qquad (3.10)$$

$$u = \frac{\log \frac{1+u_1}{1+u_0}}{\log \frac{u_1(1+u_0)}{u_0(1+u_1)}} \qquad (3.11)$$

In his book *Sequential Analysis*, Wald gives both of these equations in terms of a parameter *h*. Letting $h \to 0$ produces Equations (3.10) and (3.11). Since Wald defined $L(u)$ as the "probability of maintaining process 1," and $L(u) = 0.5$ when $h \to 0$ and $\alpha = \beta$, we will let $h \to 0$ and proceed with Equations (3.10) and (3.11). Section 4.3.1 shows that from Equation (3.10), $L(u) = 0.5$ when $\alpha = \beta$. For further discussion and definition of the parameter *h* and the operating characteristic of the test, refer to Wald, but for the purposes of our objective, we assume that $h \to 0$ since this provides our desired value of $L(u)$.

In the special case that slope *s* of the test boundaries is equal to $u/(1 + u)$, particularly when both systems have equal probability of winning the test and are therefore fairly compared, the expected number of unmatched pairs required for the test is

$$E_{\frac{s}{1-s}}(t) = \frac{-\left(\log \frac{\beta}{1-\alpha}\right)\left(\log \frac{1-\beta}{\alpha}\right)}{\log \frac{u_1(1+u_0)}{u_0(1+u_1)} \log \frac{1+u_1}{1+u_0}}. \qquad (3.12)$$

To calculate the total expected sample size required, we divide Equation (3.9) or (3.12), whichever applies, by $p_1(1 - p_2) + p_2(1 - p_1)$. Therefore, the total expected sample size for this special case is seen in Equation (3.13) [19:115].

$$E_{\frac{s}{1-s}}(total) = \frac{-\left(\log\frac{\beta}{1-\alpha}\right)\left(\log\frac{1-\beta}{\alpha}\right)}{\log\frac{u_1(1+u_0)}{u_0(1+u_1)}\log\frac{1+u_1}{1+u_0}}\left(p_1(1-p_2)+p_2(1-p_1)\right)^{-1} \qquad (3.13)$$

More investigation of expected sample sizes appears in Chapter 4, in which we perform sensitivity experiments on both the Wald $|p_1 - p_2|$ test and the Wald MSRB test and develop a methodology for implementation.

IV. Implementing the Wald Test

4.1 Introduction

We chose to measure the performance of ATR systems by their probability of identification p_{ID}, where we estimate p_{ID} by averaging the 0's and 1's which denote incorrect and correct identifications of targets drawn from the T, A, D experimental region. The identification of each target processed on an ATR system, either correct or incorrect, is thus considered a binary response. To choose the best of a set of systems which output binary responses, we usually require large sample sizes to get any clear distinction in performance if the difference in the system probabilities of success is relatively small. Wald's test for comparing two proportions can greatly reduce test sample sizes, which is particularly helpful in cases which involve destructive testing (such as testing a batch of fuses) or long processing times. In the ATR case, new systems such as MSTAR take anywhere from a few minutes to hours to process an image and determine a correct or incorrect identification of a target. The following methodology discusses the implementation of Wald's procedure to compare two or four generic competing systems, with examples from the competing configurations of the MSTAR system. The procedure can be used to compare any set of systems which output binary response data, including ATR systems for a military application. In the ensuing discussions, "configuration" and "system" will be used interchangeably, since the MSTAR configuration comparison is an example of a system comparison. In many cases, engineers compare different systems, which are not necessarily different configurations of the same system.

Throughout the following discussions, we will refer to the difference in probability of identification for two systems $|p_1 - p_2|$ which we want to detect with our test as sensitivity. WL has specified that in the comparison of two MSTAR configurations, they want to choose the superior system with 95% certainty when the p_{ID} of one system is 0.03 greater than that of the other, but are indifferent if the difference in system performance is less than 0.03; therefore, they want a test sensitivity of 0.03. Also, we will use test performance as a metric of parameter suitability to attain a certain goal (for example, to provide a certain sensitivity.) Test performance is based on an average of the scores from running the Wald test with a certain parameter set on simulated data with known results, and comparing the actual test results to the known expected results. If a test chose the known superior system as the better one, the test performance score for that run was 1; if the test chose the wrong system, the test performance score was zero. We calculate the overall test performance for that set of parameters as the sum of the test performance scores divided by the number of scores, so that if a test correctly picked the best system in 87 of 100 runs, the test performance was 0.87. Mathematically, where n is the total number of runs and we are comparing parameter settings represented by tests j:

$$z_{ij} = \begin{cases} 0, & \text{if test } j \text{ picks incorrect system as superior on run } i \\ 1, & \text{if test } j \text{ picks correct system on run } i \end{cases}$$

$$tp_j = \frac{\sum_{i=1}^{n} z_{ij}}{n} \; \forall j, \; 1 < j < \text{total number of tests} \tag{4.1}$$

As discussed above, α and β are the upper bounds on the error probabilities; now let the realized probability of incorrectly choosing system 1 for a given set of u-values be α_u, and the realized probability of incorrectly choosing system 2 for a given set of u-values, β_u.

We begin by describing two sampling schemes in Section 4.2. In Section 4.3, we explore the parameters of the Wald test and create a methodology for setting up the two-system comparison test. Section 4.4 includes a discussion of the modifications required to compare multiple systems, followed by the outline of a four-system comparison procedure. In Section 4.5, we address cautions and concerns for implementing the Wald test. Throughout the following procedure, the engineer must set and remember the experimental goals, in terms of desired sensitivity, desired α and β, sample size limit, and fairness of the sampling plan.

4.2 Sampling Schemes

Before discussing parameter choices for constructing a fair comparison of two or more systems, we will address the data handling required to execute the tests. We can collect data for the Wald test by following one of two sampling schemes. Scheme I provides the fairest comparison from the producer's point of view, while Scheme II may be the consumer's preferred method. Both use random drawn ROI's and are correct for implementing the Wald test, but the schemes account for the effects of T, A, D and E differently.

4.2.1 Scheme I: "Lock-step." In this scheme, each sample point is one randomly drawn ROI which we input to all of the systems. For example, the thirty-second sample for systems A, B, C, and D may be a T72 at 123° aspect and 42° depression. Since all systems are tested on the same ROI simultaneously, we are essentially blocking on the effects of T, A, D and E. We thus treat the performance function for system p_{ID} as follows:

$$p_{ID} = f(\text{system, T, A, D, E})$$

The "lock-step" pairing of samples acknowledges that p_{ID} for a system changes with T, A, and D, and aims to minimize the effect of those factors on the system selection. E accounts only for randomness in the observations at each T, A, D scenario, and we do not treat it as a factor in p_{ID}, but lock-step comparison blocks on E as well since all systems must process the same E at each step.

The hypothesis we test with this scheme is

$$H_o\text{: } p_1^{\,1} - p_2^{\,1} = p_1^{\,1} - p_3^{\,1} = \ldots = p_{n-1}^{\,m} - p_n^{\,m} = 0,$$

for the multiple pairwise comparison of n systems processing m ROI's. The alternate hypothesis is that at least one $p_i^k - p_j^k \neq 0$ for any two systems i and j on an ROI k, $i,j \in \{1...n\}$, $i \neq j$, $k \in \{1...m\}$.

The producers of competing systems would appreciate this sampling scheme, since every system is tested on performance on the exact same imagery in the same order, so the factors T, A, D and E do not adversely affect which system is chosen as the best by the Wald test. Systems may have strong and weak points, however, which may affect the decision. Suppose that we are testing a system which is better with trucks than tanks against one that successfully identifies tanks more frequently than it identifies trucks. If our random sequence of ROI's begins with five trucks, the producer of the system which performs better on tanks would be unhappy. We simply rely on randomness and our choice of α to provide fairness in the sampling sequence.

4.2.2 Scheme II: Random Draw. This scheme ignores the fact that the p_{ID}'s change with T, A and D, and aims to test the overall p_{ID} for each system against all others. For each sample point, we randomly draw an ROI for each of the systems. We treat performance function for system p_{ID} as follows:

$$p_{ID} = f(\text{system, all other factors})$$

Since we this scheme does not block on the other factors, which include T, A, D, and E, these other factors cause randomness in the observations of the system performance for different ROI's. We essentially treat the observed binary responses from all of the ROI's as Bernoulli trials from the same binomial population with an unknown system p_{ID}. As in Scheme I, the sampling order can cause problems depending on system strong and weak points, but in Scheme II, one system can also be tested on six fairly easy targets while a competitor receives six difficult targets. Again, randomness and a well-chosen α should minimize these problems.

The hypothesis for Scheme II is

$$H_o: p_1 = p_2 = \ldots = p_n,$$

again for n systems. The alternate hypothesis is that $p_1 \neq p_2$ for any two systems i and j, $i,j \in \{1\ldots n\}$, $i \neq j$. Note that in Scheme I, we test for differences in system performance from ROI to ROI, whereas in Scheme II, we directly compare the overall system p_{ID}'s.

Before executing any Wald test, the engineer must consider their experimental goals, and choose the sampling scheme which best suits their needs. The lock-step procedure has the advantage of requiring only one set of data collection instead of one set for each system, so it may be less expensive and therefore a better choice if the experimenter has no preference of sampling schemes according to his or her goals. However, the random-draw scheme may require slightly fewer samples, and since it requires independent sampling of the p-values, we can validly perform the ranking and selection procedure on the data if the Wald test reaches the sample size required for ranking and selection before terminating. Test parameters can be selected independently of the sampling scheme, since the parameters and the sampling scheme reflect different goals.

4.3 Setting the Wald Test Parameters

4.3.1 Selection of parameters without regard for sample size. As discussed in Chapter 3, we must begin implementation of the Wald test design by setting parameters u_0 and u_1, henceforth also referred to as the u-values. Our objective is to construct a sequential test which compares two systems equally, i.e. each system has a chance of being selected of 0.5 when $p_1 = p_2$. In manufacturing cases like the one mentioned in Section 3.3.1, we may want to favor one system, or we may want to weight the probabilities of selecting each system based on cost. However, in the ATR and MSTAR case, we assume that system cost is a negligible factor in system selection and that we simply want to choose the system with the best performance.

To compare the systems fairly, we assert that $u_0 = 1/u_1$, as mentioned in Section 3.3.1. Note in Equations (3.6) and (3.7) that only u-values, and not the parameters α and β, are used to calculate the slope of the test boundaries. Assume that $\alpha = \beta$, which will be discussed in more detail later. Given that $u_0 = 1/u_1$ and $\alpha = \beta$, the slope of the test boundaries becomes 0.5 as proven in Appendix C, and the chance of choosing either system becomes equal when $u = 1$. The likelihood function of the test, in which $L(u)$ is the probability of retaining process or system 1, is as follows [19:114]:

$$L(u)=\frac{\log\frac{1-\beta}{\alpha}}{\log\frac{1-\beta}{\alpha}-\log\frac{\beta}{1-\alpha}} \tag{4.2}$$

$$L(u)=\frac{\log\frac{1-0.05}{0.05}}{\log\frac{1-0.05}{0.05}-\log\frac{0.05}{1-0.05}}=\frac{\log 19}{\log 19-\log\frac{1}{19}}=0.5$$

Note that $L(u) = s = 0.5$, where s is the slope of the test boundaries, allowing us to use the special case for sample size calculation mentioned in Section 3.3.2.

As discussed in Section 3.3.1, α is the upper bound on the probability of incorrectly rejecting system 1, and β is the upper bound on the probability of incorrectly maintaining system 1. Assuming that all tests conclude that one of the systems is superior, α is also the upper bound on the probability of incorrectly choosing process 2, and β is also the upper bound on the probability of incorrectly failing to select process 2. If $\alpha < \beta$, then the upper bound on the probability of incorrectly choosing process 2 over process 1 is smaller than the upper bound on the probability of incorrectly choosing process 1 over process 2. In other words, the maximum probability of incorrectly maintaining process 1 is greater than the maximum probability of incorrectly removing process 1 and installing process 2 in its place. However, for the fairest system comparison, the systems must have equal opportunity to demonstrate superiority. Therefore, we will set $\alpha = \beta$. WL has requested that $\alpha = 0.05$, so we will henceforth let $\alpha = \beta = 0.05$.

To graphically demonstrate the effect of changing α and β such that $\alpha < \beta$, we refer to Figure 4.1. As seen in Equations (3.6) and (3.7), these parameters affect only the test boundary intercepts and not the slopes, so assume constant u-values. In this example, the dashed boundaries represent a test in which $\alpha = \beta$, while $\alpha < \beta$ for the test with the solid boundaries. The different intercepts can give one system a sort of head start over the other; the higher solid-line intercepts would allow the same plot of t_1 versus $t_1 + t_2$ to break its lower boundary more easily than its upper boundary. In Figure 4.1, the plot has broken the lower boundary for the test with the solid boundaries, but has not yet concluded and could still potentially break the upper boundary of the test with the dashed boundaries. Therefore, the solid boundaries allow a greater chance for the plot to break the lower boundary and choose system 1 than the dashed boundaries.

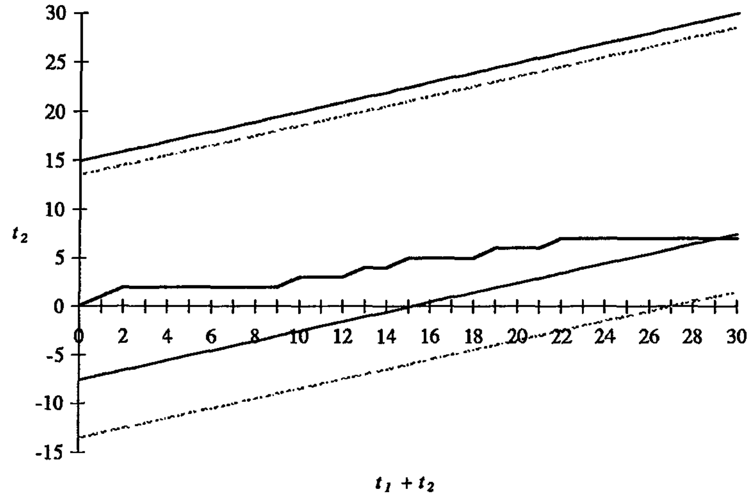

Figure 4.1. Wald boundary comparison for different (α,β) settings

dashed bounds: $\alpha = \beta$; solid bounds, $\alpha < \beta$

Given that $\alpha = \beta$ and $u_0 = 1/u_1$ to construct a test to compare two systems fairly, to find the proper u-values for an experiment, we need to start by selecting a test sensitivity level. WL expressed interest in detecting a difference of 0.03 between the p_{ID}'s of two configurations, but we will examine other sensitivities as well, since other system comparisons may require more sensitivity, and we may not have enough MSTAR data to detect the specified sensitivity. For MSTAR, the specification for overall system p_{ID} is 0.7, so u_0 and u_1 should be set such that the test correctly selects a configuration which has a p_{ID} 0.03 or more above the p_{ID} of a competing configuration, near the value of 0.7. The test should be able to detect the difference between configurations A and B when the true p_{ID}'s are at the following paired levels, which allow for 0.05 deviation above and below the performance specification, as in Table 4.1.

Table 4.1

Pairs of p_1 and p_2 with 0.03 separation around 0.7

A	0.75	0.74	0.73	0.72	0.71	0.70	0.69	0.68
B	0.72	0.71	0.70	0.69	0.68	0.67	0.66	0.65

Similar p-value pairs appear in Appendix E for other sensitivities.

To find the best u-values for sequentially testing MSTAR system configurations, we conducted simulations with the following procedure, and examined the results. Note that a similar simulation procedure can be used to find appropriate u-values if α must be different for β for some reason, such as customer specification, but a fair comparison is still desired. Appendix D describes such a procedure, though setting $\alpha = \beta$ is easier and more practical.

1. Choose the first (p_1,p_2) pair shown in Table 4.1, (0.75,0.72). Let $(\alpha, \beta) = (0.05,0.05)$. Let $(u_0, u_1) = (0.5,2)$ as a starting setting. (For MSTAR, we determined from Table 3.5 that we should investigate around $(u_0, u_1) = (0.86,1.16)$, but starting with other parameters further from 1 demonstrates the effects of different u-values on test sensitivity.)
2. Randomly generate 1000 Bernoulli trials for $p_1 = 0.75$ and 1000 for $p_2 = 0.72$, using the `binornd` command in the MATLAB® Statistics Toolbox.
3. Run the Wald test on the simulated data using a MATLAB® program.
4. Record whether the test correctly chose the system 1 (or the system with $p = 0.75$) as superior.
5. Starting at Step 1, repeat the simulation 99 more times, and average the 100 results to obtain a percentage of runs for which the test correctly chose the best system. Also record percentage of incorrect and inconclusive tests.
6. Repeat Steps 1-5 with all other (p_1, p_2) pairs shown in Table 4.1.
7. Reverse the system order and repeat Steps 1-6, such that the first (p_1, p_2) pair is (0.72, 0.75). This ensures that correct testing does not depend on system order. Average these results with those obtained in Step 5 for each (p_1, p_2) pair.
8. Repeat Steps 2-7 for other sensitivities featured in Appendix E.
9. Repeat Steps 2-8 for a range of u-values.

We began with three u_0 levels: 0.5, 0.67, and 0.75, shown in Figure 4.2. Corresponding u_1 levels are 2, 1.5, and 1.33. The y-axis represents test performance.

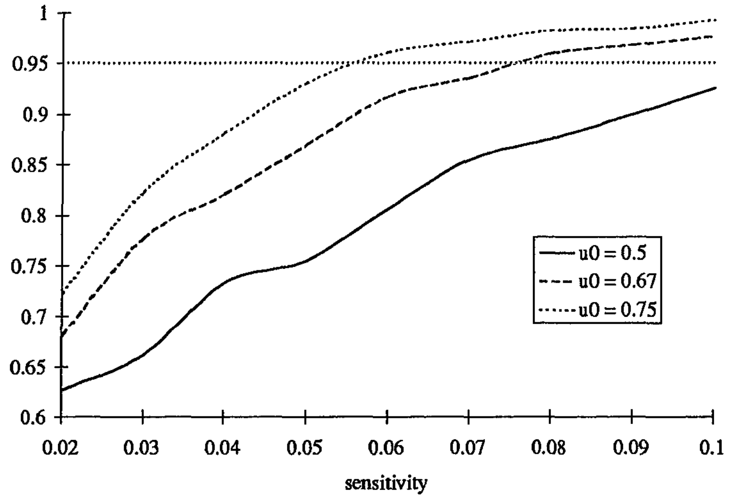

Figure 4.2. Sensitivity of the Wald test with $\alpha = \beta = 0.05$ for three u-value pairs.

Apparently, we can achieve sensitivity between 0.05 and 0.06 with $(u_0,u_1) = (0.75,1.33)$. In the simulation, this required an average of 200 samples, though Figure 4.3 shows that the average required samples varies considerably. Smoothing the data logarithmically using the MS Excel® trendline command results in the gray curves.

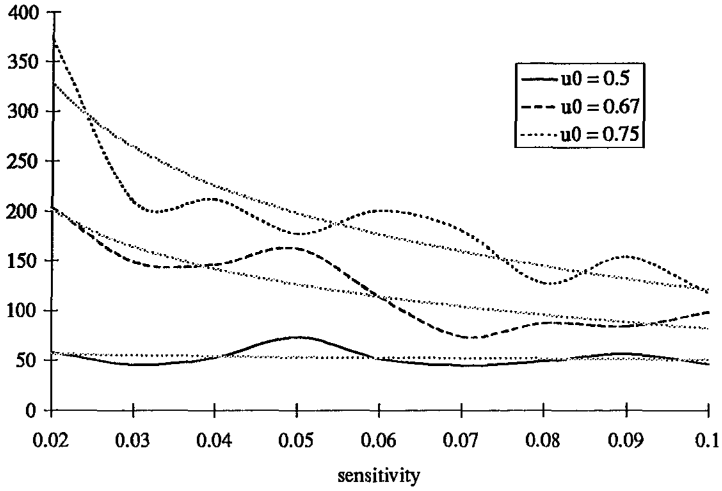

Figure 4.3. Average sample sizes required to construct Figure 4.2.

The results displayed in Figure 4.2 essentially confirm Table 3.5 for $\alpha = \beta = 0.05$. This simulation procedure therefore has validated the Wald test, since the results show that we can achieve given sensitivities with (1 - α)% accuracy when we use the u-values which correspond to the sensitivities in Table 3.5. For example, Figure 4.2 shows that when $u_0 = 0.67$, and therefore $u_1 = 1/u_0 = 1.5$, we can achieve slightly better than 0.08 sensitivity; when we refer to Table 3.5, we see that these u-values lay appear near the lower left and upper right corners, and correspond to just worse than 0.08 sensitivity (u_0 = 0.68 appears at $(p_1,p_2) = (0.66,0.74)$, and 0.74 - 0.66 = 0.08 sensitivity). The slight but favorable discrepancy on either side of 0.08 exemplifies an α_u value of less than α, since the simulation achieved the expected slightly-over-0.08 sensitivity at $(u_0,u_1) = (0.67,1.5)$ with about 97% accuracy instead of just 95%.

4.3.2. Expected sample sizes. The sample size results from the simulation plotted in Figure 4.3 demonstrate that better sensitivity requires exponentially increasing sample sizes. Though we can easily select u-values from Table 3.5 to detect the desired sensitivity and perform the Wald test, we may need more samples than we can obtain with our resources.

In his discussion of the test for $|p_1 - p_2|$, Wald gives Equation (3.12) for the expected number of (0,1) and (1,0), or unmatched, pairs required for test completion when the slope of the test boundaries s is equal to the desired likelihood $L(u)$ of the test. Since we want $L(u) = 0.5$ and $s = 0.5$ so that each test has a probability of winning of 0.5, we can apply this special case. (Recall that $L(u) = f(\alpha,\beta)$, and we let $\alpha = \beta = 0.05$ so that $L(u) = 0.5$.) Appendix C demonstrates that $u = 1$ and $s = 0.5$ in the MSTAR tests.

By inserting parameters $(\alpha,\beta) = (0.05,0.05)$ and a range of $(u_0,u_1) = (u_0,1/u_0)$ pairs into Equation (3.12), we obtain Figure 4.4. Note that as the u-values approach 1, the expected number of unmatched pairs required to complete the test increases exponentially.

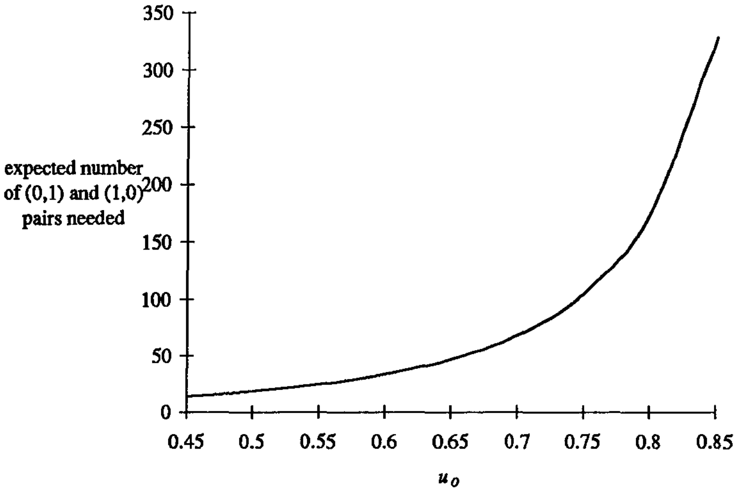

Figure 4.4. Expected number of unmatched pairs needed when $(\alpha,\beta) = (0.05,0.05)$

To find the total expected number of data points required, including matched and unmatched pairs, we refer to Equation (3.13). This equation provides an average number of ROI's required by the Wald test to detect the desired sensitivity with the selected α.

To compare the expected sample size for the Wald test to that of confidence intervals and ranking and selection, we begin with Equation (3.13) for the Wald test, and rearrange Equation (3.3) to obtain the required sample size for difference intervals in Equation (4.3), since difference intervals require smaller sample sizes than comparing independent interval estimates of each p_{ID}. We can calculate sample sizes for the ranking and selection procedure directly from tables as described in Section 3.2. We acknowledge that, as discussed earlier, binomial confidence intervals calculated for ATR systems violate the constant variance assumption in the lock-step sampling scheme, since p_{ID} is not necessarily constant for each sample. In the random-draw sampling scheme, we can use binomial confidence intervals since we sample independently and treat the samples as draws from the overall population with p_{ID} as its parameter. The lock-step scheme also violates the independent sampling assumption of the ranking and selection procedure, but the random-draw scheme does not and is thus valid. Despite the violation of independence in the lock-step sampling scheme, we will use the confidence interval and ranking and selection technique sample sizes as benchmarks to compare to the Wald test expected sample sizes, and thus assume that we are using the random-draw sample scheme throughout the sample size comparisons. Equation (4.3) provides the required sample sizes for difference intervals, which can be estimated by inserting p-values with a difference equal to the sensitivity on either side of the reference point, such as $(p_1,p_2) =$ (0.685,0.715) for 0.03 sensitivity and a reference point of 0.7.

$$n = \left(z_{1-\alpha/2} \frac{\sqrt{p_1(1-p_1)+p_2(1-p_2)}}{(p_1 - p_2)} \right)^2 \qquad (4.3)$$

The ranking and selection procedure requires smaller sample sizes than the difference intervals. In the multiple comparisons case, this occurs since ranking and selection only identifies the best binomial population, and does not attempt to rank the other populations with statistical significance as in the case of multiple pairwise difference intervals. The smaller ranking and selection sample sizes provide a more challenging benchmark for comparison to the Wald test sample sizes than the confidence intervals. These sample sizes also provide a stopping condition with fewer required samples than the confidence intervals; in other words, if a Wald test has not terminated but has collected the number of samples required to choose the best system with the ranking and selection procedure, we may simply terminate the Wald test and perform the ranking and selection procedure to find a conclusion with the minimum number of samples.

We will now examine expected sample size for the Wald test. Table 4.2 demonstrates the robustness of expected total sample size to the true difference $|p_1 - p_2|$. To construct the table, we solved Equation (3.13) with u-values for sensitivities ranging from 0.01 to 0.12, and with p-values on either side of 0.07 which reflect true $|p_1 - p_2|$ values ranging from 0.01 to 0.12. The u-values were calculated for the sensitivities via Equation (3.5), again using p-values on either side of 0.07 to reflect the desired sensitivity. For example, for sensitivity of 0.05, we used $(p_1,p_2) = (0.7 - 0.05/2, 0.7 + 0.05/2) = (0.675,0.725)$, and then reversed them to (0.725,0.675) to calculate the reciprocal u-value. The values of p_1 and p_2 are interchangeable in Equation (3.13), so we need not compensate for system order.

Table 4.2

Expected sample sizes for the Wald test

		Sens.	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12
		u0	0.95	0.91	0.87	0.83	0.79	0.75	0.72	0.68	0.65	0.62	0.59	0.56
True	*True*	*ul*	1.05	1.10	1.15	1.21	1.27	1.33	1.40	1.47	1.54	1.62	1.70	1.78
|pl-p2|	*pl*	*p2*	*Expected Sample Size*											
0.01	0.695	0.705	9101	2274	1010	568	363	252	184	141	111	90	74	62
0.02	0.690	0.710	9098	2273	1010	567	363	251	184	141	111	90	74	62
0.03	0.685	0.715	9092	2272	1009	567	362	251	184	141	111	90	74	62
0.04	0.680	0.720	9085	2270	1008	567	362	251	184	141	111	90	74	62
0.05	0.675	0.725	9075	2268	1007	566	362	251	184	141	111	89	74	62
0.06	0.670	0.730	9063	2265	1006	565	361	251	184	140	111	89	74	62
0.07	0.665	0.735	9049	2261	1004	564	361	250	183	140	110	89	74	62
0.08	0.660	0.740	9033	2257	1003	563	360	250	183	140	110	89	73	61
0.09	0.655	0.745	9015	2253	1001	562	359	249	183	140	110	89	73	61
0.10	0.650	0.750	8995	2248	998	561	359	249	182	139	110	89	73	61
0.11	0.645	0.755	8973	2242	996	560	358	248	182	139	110	88	73	61
0.12	0.640	0.760	8949	2236	993	558	357	247	181	139	109	88	73	61

Unlike the Wald test, the sample sizes for the confidence intervals are determined by the desired sensitivity and α and the p_{ID} reference point, as is evident from Equation (4.3) and explained above. The sample sizes in Figure 4.5 are the average sample sizes for the eight points at each sensitivity level given in Appendix E, as used in the simulation above. Actual results which show the change in sample size as *p*-values vary appear in Appendix F.

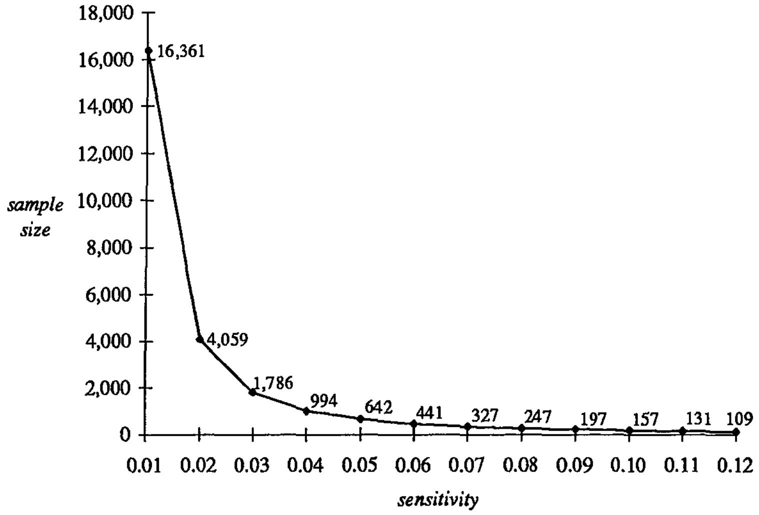

Figure 4.5. Required sample sizes for confidence intervals when $\alpha = 0.05$.

To compare the sample sizes for confidence intervals with those for ranking and selection and the Wald test, we can add the graphs of ranking and selection minimum sample sizes and Wald test expected sample sizes to Figure 4.5, omitting sensitivity levels 0.01 and 0.02 to make the graph easier to read.

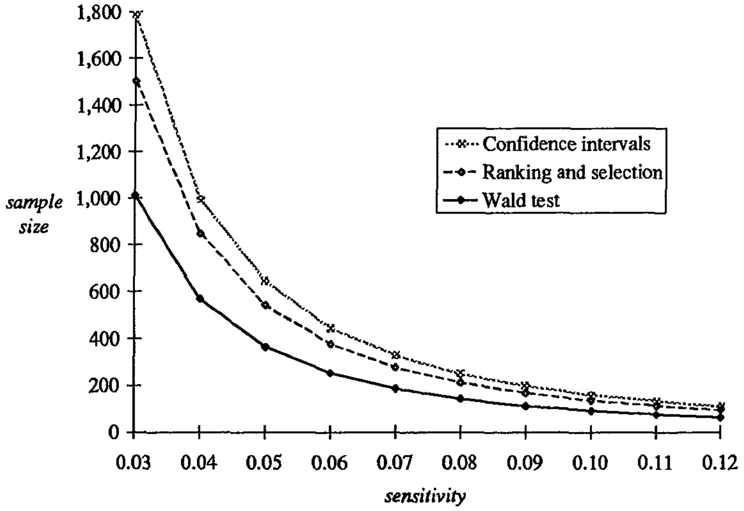

Figure 4.6. Expected sample size comparison for Wald test vs. ranking and selection and confidence intervals

Based on expected sample sizes, the Wald test requires about an average of about half as many samples than confidence intervals for $\alpha = 0.05$, and about two thirds as many samples as ranking and selection.

Since α requirements depend on the experiment, examining the relationship between sample sizes and α is worthwhile. Figure 4.7 summarizes sample sizes for the Wald test, ranking and selection (R&S), and confidence intervals (CI's) for sensitivities of 0.03 or greater. Again, we omit 0.01 and 0.02 for legibility, and the order of the key corresponds with the order of the lines as they appear on the chart. Note that the Wald test is designated by solid markers, the R&S procedure markers are hollow, and the CI markers are gray with a black border. Also notice that the expected sample size for the Wald test at $\alpha = 0.05$ is smaller than that for confidence intervals at α =0.10. We will use

R&S sample sizes as our stopping condition; CI sizes are included merely for demonstrative purposes.

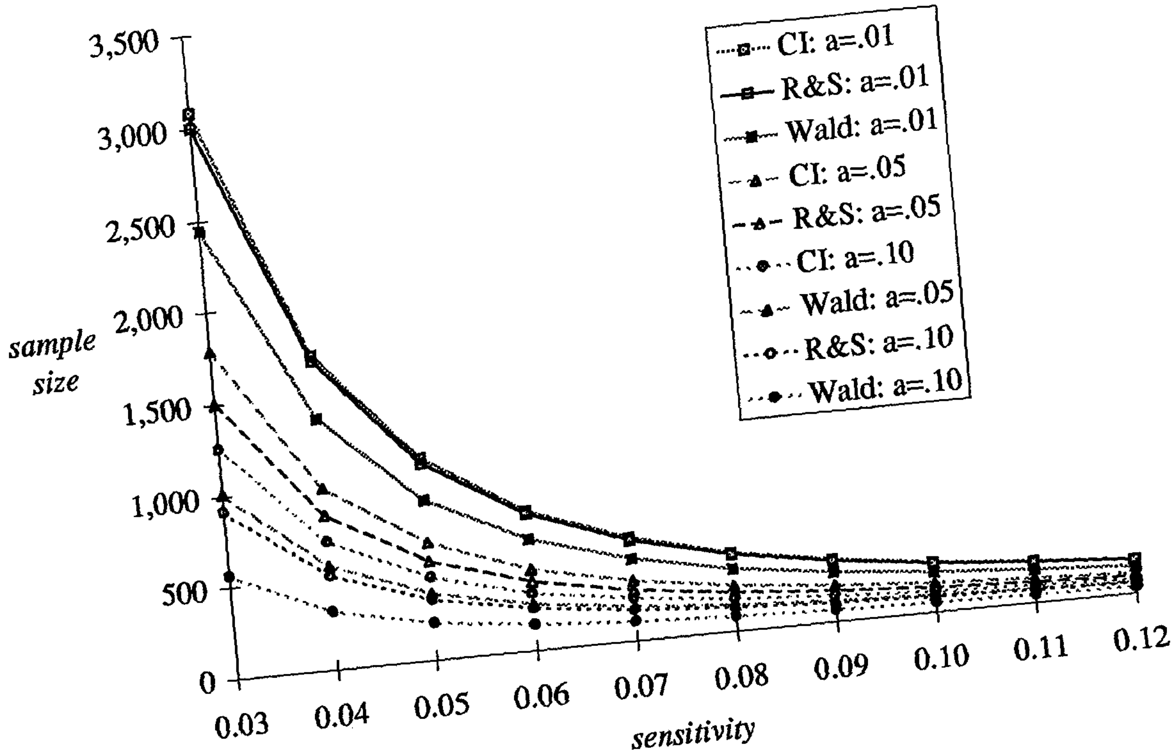

Figure 4.7. Sample sizes for different α values

Section 4.3.3 will discuss use of these chart. To examine sample size requirements for sensitivities of 0.06 or greater, we could reconstruct this chart with a different range for better resolution. The chart can be easily reconstructed for use in other experiments, but we will focus on ranges useful for the MSTAR test.

When selecting parameters, we must keep in mind that Equation (3.13) provides expected sample sizes, which may not reflect the actual experimental requirements very well. Using the confidence interval sample size as a stopping condition guarantees that the experiment cannot require a larger sample size than the precalculated confidence interval sample size, but we can do much better if the Wald test terminates close to its expected sample size.

4.3.3. Parameter selection procedure. Given the above examination of the behavior of the parameters, the following procedure provides instructions for setting up the Wald test to fairly compare two systems.

1. Determine whether a sample size constraint exists, due to time or resources.
2. Identify your desired test sensitivity, desired α, and reference point for the p-values. Construct a figure similar to Figure 4.7 using Equation (4.3) and Equation (3.13) for the range of p-values near your reference point, using the methods discussed to construct the figures above.
3. Refer to Figure 4.7. Choose the best sensitivity level and α which you can achieve with the sample size constraint. This will require balancing the α and sensitivity requirements as well. Find the sample size required to detect this sensitivity and α with ranking and selection for the difference in the two proportions.
4. Calculate the u-values corresponding to your achievable sensitivity using Equation (3.5). Refer to the construction of Table 3.5 as an example.
5. Let $\beta = \alpha$. Also verify that $u_1 = 1/u_0$, using u-values from Step 4.

6. Perform the Wald test. Stop when the test terminates or when the sample size reaches that required for ranking and selection. In the latter case, perform the procedure and draw a conclusion.

For example, if we can afford only 500 images to test MSTAR and absolutely require α = 0.05, then according to Figure 4.7, our sensitivity will be approximately 0.045. Table 3.5 shows that u-values of (0.8,1.25) should work; note that 0.8 = 1/1.25. We can check this by calculating Equation (3.5) for (p_1,p_2) = (0.7 - 0.045/2, 0.7 + 0.045/2) = (0.6775,0.7225), so the u-values are (0.81,1.24) when the true p-values surround 0.70. If we want to achieve 0.03 sensitivity with α = 0.05 as in the MSTAR case, we need to obtain 1000 samples, according to Figure 4.7. We can calculate this exactly as well, but since this number is only an expected value and not an exact deterministic sample size requirement, reading the graph provides sufficient information for planning purposes. We can expect to need about 1000 samples, and should decide if this is feasible given our resources. Our R&S sample size is 1502.

Figure 4.8 summarizes the procedure described above.

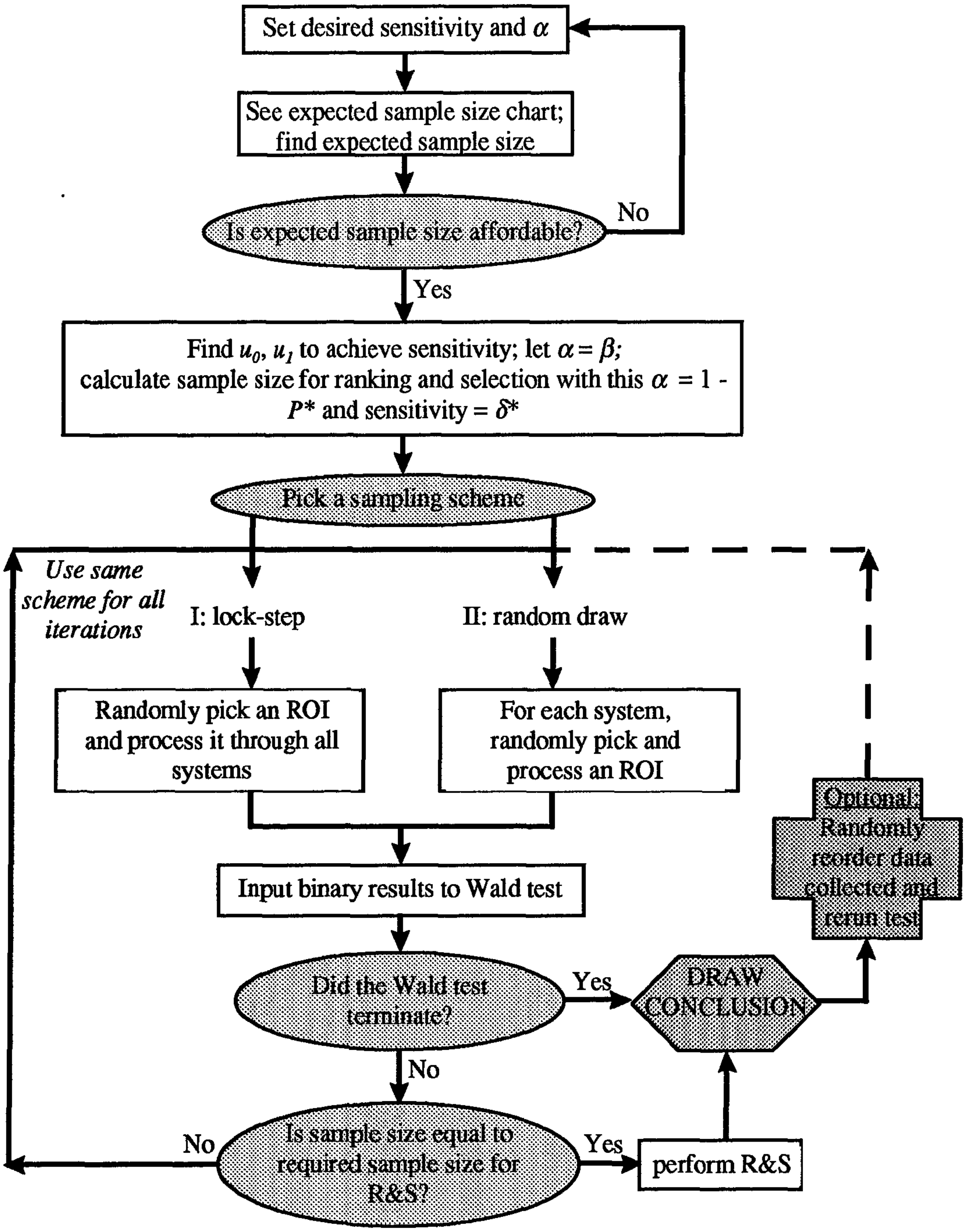
Set desired sensitivity and α
See expected sample size chart; find expected sample size
Is expected sample size affordable?
No
Yes
Find u_0, u_1 to achieve sensitivity; let $\alpha = \beta$; calculate sample size for ranking and selection with this α = 1 - P^* and sensitivity = δ^*
Pick a sampling scheme
Use same scheme for all iterations
I: lock-step
II: random draw
Randomly pick an ROI and process it through all systems
For each system, randomly pick and process an ROI
Input binary results to Wald test
Did the Wald test terminate?
Yes
DRAW CONCLUSION
Optional: Randomly reorder data collected and rerun test
No
Is sample size equal to required sample size for R&S?
No
Yes
perform R&S

Figure 4.8. Wald test procedure for fair comparison of two systems.

To summarize, balancing of the sample size, α, and sensitivity constraints depends on the experiment, and appropriate tradeoffs should be made according to the experimenter's objectives and required accuracy. The expected sample sizes are only estimates of the actual required sample sizes, so allowance should be made for actual required sample sizes larger than expected. For example, if we absolutely cannot exceed 1000 samples in an experiment, we should pick a sensitivity and/or α which corresponds to an expected sample size of about 800. Engineering judgment thus dictates parameter settings.

4.4 Comparison of Four Systems

To fairly compare four ATR systems, we introduce the MSRB procedure into the Wald test to create what we will call the Wald MSRB test. The MSRB procedure states that the significance level α of each pairwise test is the desired overall significance level α_0 divided by the number of pairwise comparisons. In the four-system test, we begin with $\binom{4}{2} = 6$ comparisons and $\alpha = \alpha_0/6$, then drop to $\binom{3}{2} = 3$ comparisons with $\alpha = \alpha_0/3$, and then $\binom{2}{2} = 1$ comparison with $\alpha = \alpha_0/1$ as the test eliminates each system. As the value of α changes throughout the test, the test boundaries move closer together each time a system is eliminated, as prescribed by Equations (3.6) and (3.7). Figure 4.9 exemplifies a

Wald MSRB test with constant u_0 and u_1, where the center trace displays the sampled pair data plot, and the outer lines are the test boundaries.

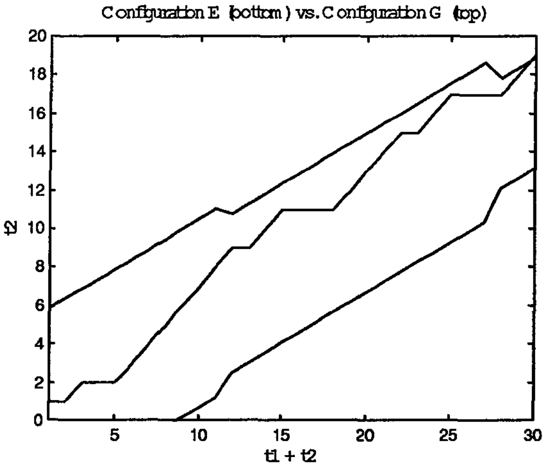

Figure 4.9. MSRB Wald test for multiple comparison of systems (t_1 + t_2 counts unmatched pairs (0,1) and (1,0); t_2 counts only (0,1) pairs)

4.4.1 Test parameter investigation. As in the two-system (henceforth two-system) comparison test, we must begin by selecting the *u*-values. As discussed in Chapter 3, these values represent efficiency ratios of two systems. In the four-system case, we are simply performing simultaneous six pairwise comparisons, so the *u*-values chosen in the single pairwise test apply. We begin with (u_0,u_1) = (0.867,1.153) for all three tests, and obtain the boundaries shown in Figure 4.10.

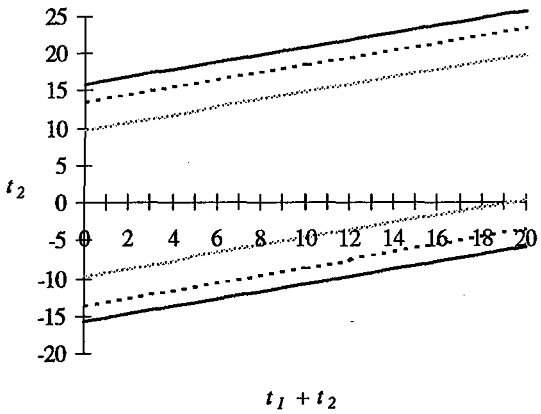

Figure 4.10. Starting Wald boundaries for multi-system comparison tests.

Gray lines - final test; Dashed lines - intermediate test; Solid black lines - initial test.

The further apart the boundaries lie, the more samples we need to conclude a test. Ideally, we would like adjustable *u*-values which change each time a configuration is dropped and the significance and power of the individual tests increase, so that the boundaries for the four-system comparison exactly follow those of the single comparison test and the tests can thus conclude with a minimal sample size. Unfortunately, this would reduce our test sensitivity to a coarser level than our goal. We verified this by running the Wald test 500 times on different sets of data simulated for four hypothetical systems with p_{ID}'s (0.715,0.685,0.685,0.685) and using only the 2-system comparison boundaries with parameter vector (α,β,u_0,u_1) = (0.05,0.05,0.867,1.153), which corresponds to 0.03 sensitivity. The test chose the system with p_{ID} = 0.715 as superior in only 87.7% of runs, less than the 95% required by specifications. The Wald MSRB procedure with the same *u*-values selected the p_{ID} = 0.715 system in 95.2% of 500 runs, thus achieving our test goals of 95% accuracy in detecting 0.03 sensitivity.

Though we cannot adjust the test boundaries to lower the expected sample size and maintain the desired sensitivity and/or α and β, we may want to adjust parameters to correspond to a desired sample size. In addition to examining Figure 4.7, we can investigate the exact tradeoff between α and β and the u-values, which correspond to sensitivity as shown in Table 3.5, if we decide that we need to adjust our parameters to better reflect our available sample size. Appendix G demonstrates the mathematics of this tradeoff in detail when $\alpha = \beta$ and $u_0 = 1/u_1$ for a constant expected sample size (in other words, constant test boundaries).

4.4.2 Test parameter selection procedure. The procedure for constructing the 4-system test is no more difficult than for the 2-system test. The 4-system test does allow for screening experiments for the 4- and/or 3-system comparisons by starting with a coarser sensitivity and thus a narrower region between the boundaries, but coarser initial sensitivity means that the chance of incorrect test outcomes increases, so this is not recommended. For example, in Figure 4.2, when $u_0 = 0.67$, we can achieve sensitivity of 0.075 with 95% accuracy; however, sensitivity of 0.03 is only attainable with 77% accuracy. In 23% of runs, systems with p_{ID}'s 0.03 apart will be incorrectly eliminated in the screening part of the experiment (inconclusive tests cannot occur since sampling will continue through the final 2-system selection stage). Therefore, the author recommends choosing one sensitivity and using it throughout the experiment, which is easier to implement when given a sensitivity specification anyway.

The analyst should plan sample size requirements based on the expected sample size for the initial test, which has the smallest α. Figure 4.11 displays a sample size chart for the four-system comparison when overall α =0.05.

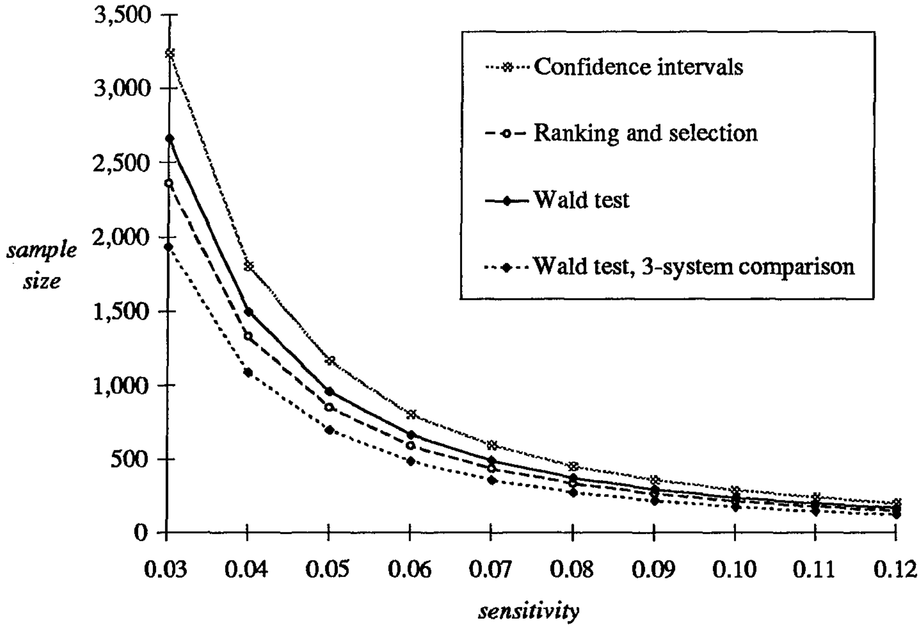

Figure 4.11. Sample sizes for the four-system test when overall $\alpha = 0.05$

Notice that the expected sample sizes for the Wald test are greater than the minimum required sample sizes for the ranking and selection procedure when we compare 4 systems. This means that on average, we will reach the stopping condition of the R&S sample size before the Wald test terminates if the difference between the highest and lowest p_{ID}'s is equal to the sensitivity. In many experiments, however, the difference between the best and worst systems will be greater than the sensitivity, so that the Wald test will usually terminate well before the expected sample size, at which point the significance level for the remaining hypotheses increases according to the MSRB procedure. The expected sample sizes for greater values of α are lower, so the Wald MSRB test may still require fewer samples than the R&S procedure. The curve of the 3-system Wald test demonstrates that after the first hypothesis is rejected by the Wald test, the expected sample size is reduced to less than the R&S sample size. In any case, by setting the R&S sample size as a stopping condition, we will find the best system with the minimum sample size from one of the tests (Wald MSRB or R&S.)

To set up the multiple-system test, execute the following instructions.

1. Determine whether a sample size constraint exists, due to time or resources.
2. Identify your reference point for the *p*-values, desired overall α for the test, and the desired test sensitivity. Construct figures similar to Figure 4.11 using Equation (4.3) and Equation (3.13) for the range of *p*-values near your reference point, using the methods discussed to construct Figure 4.7 for the two-system case. Use α/M to construct the figures, where M is the number of pairwise comparisons and α is the desired overall significance level.
3. Refer to Figure 4.11. Choose the best sensitivity level and overall α which you can achieve with the sample size constraint, keeping in mind that the expected sample size for the Wald test is probably a very conservative estimate. Again, this choice will require balancing the α and sensitivity requirements with the expected sample size. Determine the sample size required for the R&S procedure to determine the best system with this sensitivity (δ*) and α (or 1-*P**).

4. Calculate the *u*-values for your achievable sensitivity using Equation (3.5). Refer to the construction of Table 3.5 as an example.

5. Perform the Wald test. Drop configurations and hypotheses as tests terminate. Stop when only one system remains, or when the sample size reaches that required for the R&S procedure. In the latter case, perform the R&S procedure and draw a conclusion.

The MATLAB® code in Appendix H performs the Wald MSRB test, and can also perform the 2-system comparison test. The code does not account for the stopping condition, however, though this feature could be easily added.

Figure 4.12 summarizes this procedure.

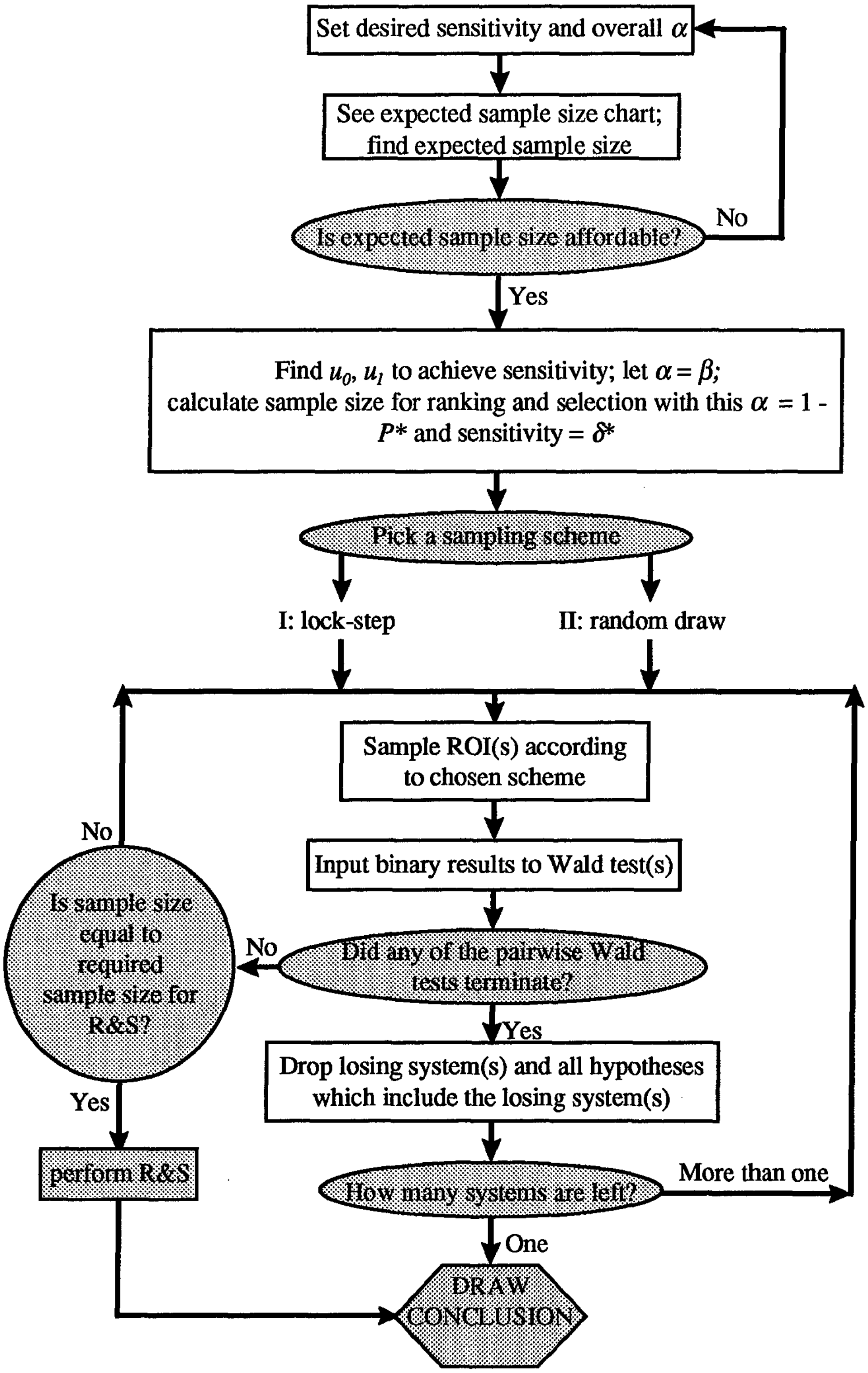
Set desired sensitivity and overall α
See expected sample size chart; find expected sample size
Is expected sample size affordable?
No
Yes
Find u_0, u_1 to achieve sensitivity; let $\alpha = \beta$; calculate sample size for ranking and selection with this α = 1 - $P*$ and sensitivity = $\delta*$
Pick a sampling scheme
I: lock-step
II: random draw
Sample ROI(s) according to chosen scheme
Input binary results to Wald test(s)
Did any of the pairwise Wald tests terminate?
No
Yes
Is sample size equal to required sample size for R&S?
No
Yes
Drop losing system(s) and all hypotheses which include the losing system(s)
How many systems are left?
More than one
One
perform R&S
DRAW CONCLUSION

Figure 4.12. Wald test procedure for fair comparison of multiple systems.

4.5 Final Comment on Sample Size Restrictions

In many experiments, the true p_{ID}'s for two compared systems may be further apart than the required sensitivity. In these cases, the tests are likely to conclude with less than the expected sample size, and we have saved ourselves considerable resources. However, sometimes the sample size allowed falls short of that needed for testing. In these cases, we generally must settle for coarser sensitivity or overall test significance, but we may also try the rerun procedure. We can also use this procedure to verify results from tests which concluded within specifications, since it requires no additional data collection, only running additional Wald tests on the computer.

We first execute the test with parameters which give the desired sensitivity, and record the result if the test concludes. We then would take the data collected for this first test, randomize the order, and rerun the Wald test several times without collecting any new data. If the test continues to choose the same superior system, we may assume that the test has chosen the true optimal system. If most of the tests fail to conclude, we may decide to collect more data, depending on engineering judgment. If the data collected on the first run is not a representative sample, we risk choosing the wrong system; also, we may collect too small a data set on the first run to effectively perform reruns, should the original test conclude quickly. In the first case, we would have a difficult time verifying whether the sample was representative, and therefore must ensure that the data points are selected truly randomly from an experimental region which is known to be representative of overall system performance. In the second case, most of the reruns will not conclude, so we would have few to no verification runs. Despite these risks, the rerun procedure is an option which can preserve the minimum-data optimality of the Wald test in many experiments. The procedure can also confirm that the test is robust to the random data order, especially when we have collected enough data to conclusively rerun the Wald test many times.

In Chapter 5, we will apply both the 2-system and 4-system comparison procedures to real and simulated data. In the real-data 2-system case, we will try both of the sampling schemes and compare the results.

V. Optimal MSTAR System Selection

In Chapter 4, we applied the Wald test to the fair comparison of two systems by setting the parameters such that $\alpha = \beta$ and $u_0 = 1/u_1$. We then outlined the procedure for setting α and u_0 to reflect experimental goals and sample size constraints, and introduced the MSRB procedure for the comparison of multiple systems. Also, we discussed two sampling schemes for executing the test. With this test methodology in hand, we can proceed to run the procedures on actual and simulated MSTAR data.

The MSTAR system developers identified several system comparisons to perform. For the FE and M module developer selection, WL required a comparison of the four system configurations analyzing peaks, followed by a comparison of the four systems analyzing regions. They also specified one feature extractor and one match module, and wanted to determine whether this configuration performed better when analyzing peaks or regions in the SAR imagery. To compare the peak and region versions of the specific FE/M system configuration, we simply applied the Wald test methodology described in Section 4.3.3. We will refer to this 2-system comparison as the P/R test. To conduct the comparisons of the four peaks systems and the four regions systems, henceforth called the peaks test and regions test respectively, we used the procedure of Section 4.4.2. We also compared our P/R results to confidence intervals like the one in Equation (3.2) to verify our test results. The results of those experiments follow a survey of the data.

5.1 The MSTAR Data

WL provided the author with ten binary data sets: one for each of the four peaks configurations, labeled A - D; one for each of the four regions configurations, labeled E - H; and two longer data sets for a particular set of FE and M modules, one analyzing peaks and one analyzing regions, labeled P and R respectively. Each set consisted of the binary processing results for ROI's which were randomly selected from a representative experimental region of all targets at all aspect angles and 30° and 45° depression angles (data was unavailable for other depression angles). Though sets P and R are each results from MSTAR configurations included in sets A - H, we kept the data separate because sets R and P were produced after developers modified the processing algorithms. Therefore, the R and P data sets resulted from processing imagery on slightly different systems than their counterparts C and G.

Summaries of the MSTAR data appear in Appendix I. Section I.1 presents two sets of confidence intervals for p_{ID} of each system. The first set of intervals, labeled "all data points," provides 95% normal confidence intervals for all of the data provided for each system. The "common data points" intervals resulted from a screening of the ROI's tested to construct the data sets, such that only ROI's tested by all of the regions, peaks, or P/R configurations remained in the data set. For example, the P/R common data set contains 962 points, while the R all points set contains 1152 and the P all points set contains 1048. This means that system R was tested on 1152 - 962 = 190 ROI's that were not introduced to system P, and system P processed 86 images not processed by system R. One may argue that the common data sets provide fairer competition between the configurations, since the common sets are results from the systems processing the exact same imagery, whereas in the all-points sets, the average level of complexity of the imagery processed by one system could be less than for another system, so that one system's p_{ID} estimate appears inflated by the effects of T, A, D and E when compared to the other. On the other hand, the all-points sets could comprise more representative samples of each system's performance, and allow more points for random draws. In most sequential testing situations, we collect data as we test and not before as in the MSTAR case, so this decision will rarely arise in the implementation of the Wald test. We therefore made the engineering judgment decision to use the common data set for the P/R test in the interest of fairness. We performed the test with both of the sampling schemes described in Section 4.2.

Though the P/R common data set was large enough to obtain useful results, because of the limited size of the system A - H data sets, none of the 500 peaks or regions tests concluded with the 141 and 244 common data points available, respectively. We therefore decided to simulate data for these tests, based on $\hat{p}_{ID}$'s from the all-points data set. We could also have used the $\hat{p}_{ID}$'s from the common data set, but since the all-points data consisted of more samples for each system, it may provide better estimates of overall system $\hat{p}_{ID}$. Also, the all-points data $\hat{p}_{ID}$'s are closer together and thus provide a more interesting demonstration of Wald test. Since we are not using real data, results from simulating data based on either set of $\hat{p}_{ID}$'s does not provide a true comparison of the peaks and regions systems, so we may choose either set for demonstrative purposes.

Figure 5.1 displays these 95% confidence intervals for the all-points data, and shows no statistically significant differences between any configurations when we compare the $\hat{p}_{ID}$ intervals. (The intervals do show that system G is better than system A, but since A is a peaks system and G is a regions system, we are not comparing them.)

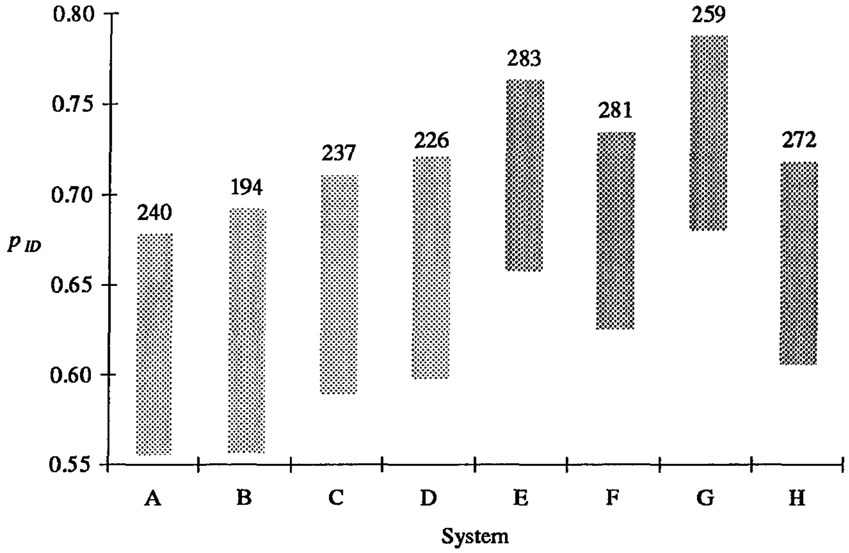

Figure 5.1. MSTAR configuration A - H data confidence intervals. Sample sizes appear above the interval bars.

Section I.2 displays difference intervals for the $\hat{p}_{ID}$'s for all of the system comparisons, nearly all of which indicate no statistically significant differences given the available data. The hypothesis for all of the intervals is H_0: $p_i = p_j$ versus H_1: $p_i \neq p_j$, for any two systems i and j. Note that for the common data sets, system R has a statistically significantly higher $\hat{p}_{ID}$ than system P, but in the all-points data, the confidence interval for $\hat{p}_R - \hat{p}_P$ does not include zero and therefore does not demonstrate a significant difference. Since we chose to use the common data set for the Wald test example, we will also use the confidence interval resulting from the common data set and conclude that $p_R > p_P$ with 95% confidence. By design, the Wald test can identify the true optimal configuration and resolve this dilemma, assuming that the common data set is a representative sample of the experimental region.

5.2 *The 2-System P/R Comparison*

We implemented the 2-system test by following the procedure in Chapter 4 and using both sampling schemes for experimental purposes, to investigate whether the sampling scheme choice makes a difference in test accuracy or sample size. For MSTAR, the schemes are applied as follows:

- Scheme I: the results for the same ROI processed with each system were paired (e.g. the twenty-second data point is the binary result pair for system R and system P both processing the exact same ROI of a T72 at some A and D)
- Scheme II: the scenarios were randomly drawn for each system (e.g. the twenty-second data point is the binary result pair for system P processing a certain ROI of a T72 at some A and D, and for system R processing an ROI of a Scud at some A and D).

Pre-collected data permitted the testing of both sampling schemes, but in most experiments, we can only test with the scheme we use to collect the data. Scheme choice depends on experimental goals, as discussed in Section 4.2, but Scheme I is generally less expensive since we need only pay for the collection of one set of images. Also, Scheme I blocks on the T, A, D, E effects and thus provides the fairest comparison. Therefore, Scheme I is the default choice when the experimenter has no preference but wishes to minimize data collection. The independent sampling of Scheme II allows for easier analysis since the R&S procedure is valid in this case, so this scheme may be preferable if data is unlimited.

According to Table 4.2 and Figure 4.7, we need about 1000 samples to achieve 0.03 sensitivity at the $\alpha = 0.05$ level, which meets WL specifications. Given just over 962 samples in the common data set, we will proceed with these parameter settings. Note that the R&S procedure requires 1502 samples to detect a difference of 0.03 at the $\alpha = 0.05$ level, so we do not have enough data to use the stopping condition.

We ran the Wald test 500 times with parameter vector (α,β,u_0,u_1) = (0.05,0.05,0.867,1.153) to detect 0.03 sensitivity. Table 5.1 contains a summary of the results.

Table 5.1

Results of the P/R test

Data set/sampling scheme	*Common data, Scheme I*	*Common data, Scheme 2*
Percent of tests which concluded with the available data	100%	100%
Percentage of completed tests which chose the system with the highest p_{ID} (system R)	100%	100%
Average number of samples required*	290	282
Standard error of required number of samples	90.1	101.7
95% normal confidence interval for the average number of samples required*	(282,298)	(273,291)
Maximum number of samples required	550	595
Minimum number of samples required	111	84

*rounded up to the next integer

Apparently, we met our objective of choosing the optimal system with 0.03 sensitivity and 95% minimum accuracy with the available data, assuming that system R is superior to system P as discussed previously. Conversely, the Wald test verifies that system R is superior to system P.

Scheme I appears to require slightly more samples but a smaller standard error than Scheme II. This makes sense, since matched pairs are more likely to arise when the two systems are testing the same ROIs at each point, and more matched pairs means a higher overall sample size is necessary to collect enough unmatched pairs for a conclusive test. For example, an M35 at 30° aspect and 30° depression may generally be a much easier target to identify than a T72 at 90° aspect and 45° depression, so if we randomly draw the M35 for testing system R and the T72 for testing system P as our thirtieth data point, we are more likely to get a (1,0) (for (R,P)) result than if both systems were tested on the M35 for that data point.

To test the hypothesis H_0: $\hat{p}_R - \hat{p}_P = 0.03$, H_1: $\hat{p}_R - \hat{p}_P \neq 0.03$ with a 95% confidence interval, we need about 1800 samples, while the Wald test only required between 282 and 298 samples for Scheme I, and between 273 and 291 samples for Scheme II, 95% of the time. The R&S procedure requires 1502 samples to test the same hypothesis, while the confidence intervals require about 1800. In both sampling schemes, the Wald test has significantly reduced the data and processing requirements to select the optimal system in a 2-system comparison as compared to the R&S procedure or the confidence interval.

Since the peaks and regions systems comparisons require about 3300 samples to construct confidence intervals to detect 0.03 difference when $\alpha = 0.05/6$ as in the initial 4-system comparison, which starts with the six pairwise hypotheses mentioned in Section 3.3.1, the peaks and regions sample sizes allow even more room for reduction. Even the R&S procedure requires 2361 samples to choose the best of four systems.

5.3 *The 4-system Peaks and Regions Comparisons*

To implement the peaks and regions tests, we followed the 4-system comparison test procedure outlined in Section 4.4.2, using the same parameter vector as for the P/R test and applying the MSRB. The expected sample size to detect 0.03 differences in p_{ID} with the Wald MSRB test is 2658 for the worst-case scenario, and since the WL data sets contained under 300 samples for each configuration, only 1 run out of 500 peaks tests and none of the 500 regions test runs concluded. (The worst-case scenario occurs when the p_{ID}'s of best and worst systems have a difference equal to the sensitivity.) Therefore, WL needs more data to choose the optimal system in each group, or needs to choose a coarser sensitivity or larger α.

Rather than performing the tests with much coarser sensitivity or a large α and very little data, we treated the all-points data $\hat{p}_{ID}$'s as the true p_{ID}'s, and used them to simulate Bernoulli trials as data. As mentioned previously, we chose the all-points $\hat{p}_{ID}$'s instead of the common data $\hat{p}_{ID}$'s for demonstrative purposes, since the all-data $\hat{p}_{ID}$'s are closer together (particularly in the peaks case) and therefore better demonstrate Wald test performance when relatively small differences exist between system $\hat{p}_{ID}$'s. Therefore, we simulated 10,000 data points for each system, to guarantee no data limit, and performed the Wald MSRB test on simulated data for both the peaks and regions systems. Since we use simulated data generated from the $\hat{p}_{ID}$'s and not from some system performance function $f(p_{ID})$ = (system, T, A, D), this test mimics Sampling Scheme II. Table 5.2 displays the results.

Table 5.2

Simulated peaks test and regions test results

Test	*Peaks*	*Regions*
Percent of tests which concluded with the available data	100%	100%
Percentage of completed tests which chose the system with the highest p_{ID}	73%	93%
Percentage of completed tests which chose the two systems with the highest p_{ID}'s**	99.2%	100%
Average number of samples required by completed tests*	1159	856
Standard error of required number of samples, completed tests	729.4	551.1
95% normal confidence interval for the average number of samples required by completed tests *	(1095,1223)	(808,905)
Maximum number of samples required by a completed run	5269	4219
Minimum number of samples required by a run	180	159

*rounded up to the next integer

**since the highest two $\hat{p}_{ID}$'s are within 0.03 of each other, and thus either of the best two peaks or regions systems is an acceptable choice.

Since we simulated the data and are treating the $\hat{p}_{ID}$'s as the true p_{ID}'s, by construction, $p_D > p_C > p_B > p_A$, and $p_G > p_E > p_F > p_H$ in this experiment. In the regions case, $p_E - p_F = 0.0306$, so the Wald test with our parameters should be able to declare systems E and/or G superior to F and H at least 95% of the time. The results show 100% success in making this distinction with about a quarter of the 3300 samples required to make the 4-system pairwise distinction with confidence intervals and just over one third of the 2361 samples needed for R&S, on average. The test even concluded that $p_G > p_E$ in 93% of cases, where $p_G - p_E = 0.0243$.

The p_{ID}'s used for the peaks test were much closer together; $p_D - p_A > 0.03$ and $p_D - p_B > 0.03$, and $p_C - p_A > 0.03$, but $p_D - p_C = 0.0095$, so under the sensitivity specification of 0.03, the test should not be able to successfully choose p_D over p_C, but should successfully eliminate systems A and B. The Wald test did this in 99.2% of the runs with an average of about half of the 2361 samples required by ranking and selection, and one third of the 3300 samples needed for confidence intervals.

In the procedures of Chapter 4, we instruct the engineer to stop sampling when the Wald test sample size met the R&S sample size. We did not use this stopping condition in the peaks and regions experiments, merely to demonstrate how large the sample size can grow, and to better model the distribution. The maximum sample sizes were quite a bit higher than the R&S sample size of 2361, but the 95% confidence intervals for the sample sizes showed that the Wald test needed only about half of the samples required by the R&S procedure. Therefore, the maximum sample sizes are well into the tails of the distributions, and using the R&S sample size as a stopping condition as discussed in Chapter 4 ensures that we cannot require more samples than the R&S procedure. The 95% confidence intervals for the sample sizes of the experiments showed that we can do considerably better with the Wald test, however, most of the time.

So far in these experiments, we have used knowledge of the true p_{ID}'s or the superior system to validate the results of the Wald test. Suppose now that we have no knowledge of the p_{ID}'s, but were only supplied with the data for the 2-system P/R and the simulated data for the 4-system peaks and regions tests. The Wald test has thus concluded the following answers to the WL system selection questions with $\alpha = 0.05$, allowing for the example that the all-points $\hat{p}_{ID}$'s used to simulate the peaks and regions data are accurate estimates of the true p_{ID}'s:

1. In the P/R test, the regions configuration is optimal
2. In the peaks test, configurations C and D are better than A and B, and p_C and p_D are within 0.03 of each other in p_{ID}
3. In the regions test, configurations E and G are better than F and H, and p_E and p_G are within 0.03 of each other in p_{ID}.

Still allowing that the all-points $\hat{p}_{ID}$'s used to simulate the peaks and regions data are accurate estimates of the true p_{ID}'s for the example, WL can base the choice between systems E and G on tiebreaking factors such as system cost, but cannot declare one or the other superior in terms of probability of target identification within the specified margin of 0.03. Assuming that the P/R test has demonstrated that regions systems are better than peaks systems in general (a conclusion based on engineering judgment), WL may decide to discard systems A through D entirely. Note that we have drawn these conclusions based on simulated data for the regions and peaks tests; if the p_{ID}'s used to generate the simulated data are reflective of the true p_{ID}'s then the conclusions are valid, but we cannot show this without more sampling. We would usually draw conclusions based on actual data; the simulated data simply provided an illustration of implementation of the test.

Figure 5.2 displays the final comparison graph of a peaks test which took 512 samples to complete. Note how the boundaries converge quickly (and discretely) as hypotheses, and thus systems, are rejected near the end of the test.

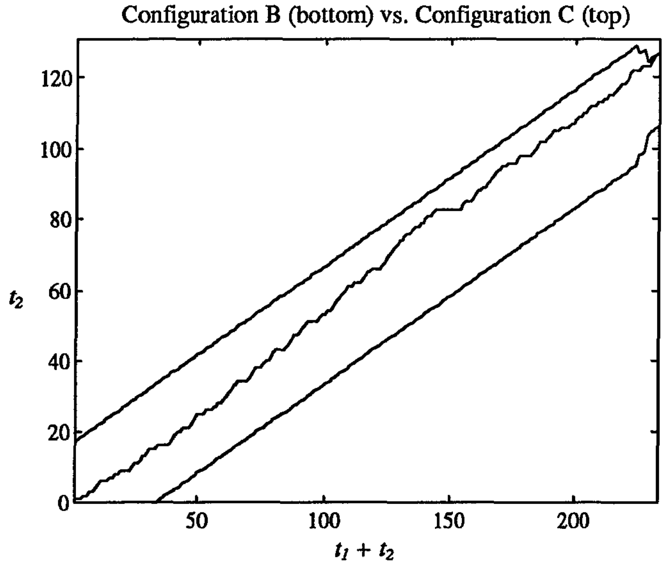

Figure 5.2. Sample Wald MSRB peaks test with $\alpha = \beta = 0.05$

Figure 5.3 displays the final comparison graph of a regions test which required 986 samples to complete. This graph clearly shows the sequential system rejections.

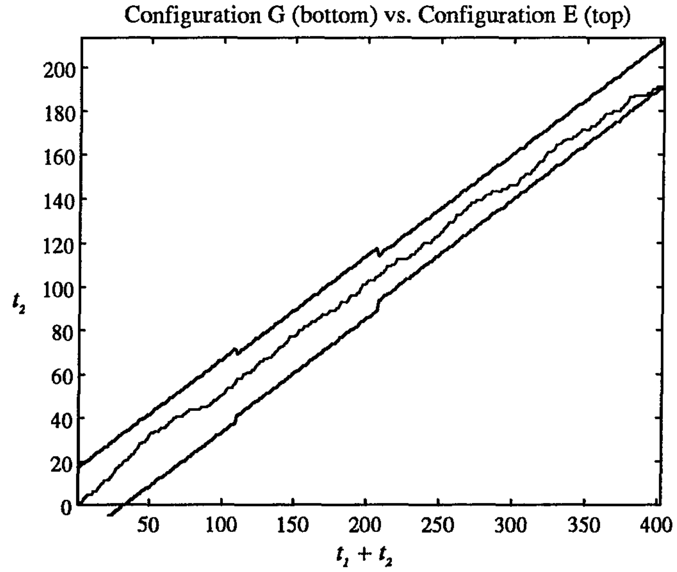

Figure 5.3 Sample Wald MSRB regions test with $\alpha = \beta = 0.05$

VI. Comments, Conclusions and Recommendations

6.1 Comments and Conclusions

The Wald test for comparing two proportions can be applied to compare any two systems which produce binary responses, and can be easily initialized to compare systems fairly based on randomly ordered data. In the ATR case, the fairest comparisons occur when each data point compares the performance of different systems on the same image, or at least the same target in the same conditions, in the case of ATR systems which use different sensors. By adding the MSRB procedure and thus changing the α and β as systems are rejected, several systems can be compared simultaneously with sequential testing. Fair comparison entails that the different systems are tested on images drawn from the same population. In comparing ATR systems which analyze different imagery types, such as SAR versus forward-looking infrared radar (FLIR), experimenters should obtain imagery of the same targets in the same conditions, angles, etc. if feasible. Conditions should also be chosen so that the systems are compared on the same standard, and not on each individual system's strong or weak points. Analysts should select a fair mix of imagery representing the spectrum of capability for each system and randomly order the images. If for some reason data of the same mix of targets and conditions is unavailable for two systems which use different sensors, different randomly drawn images for each system may be used in the test, but a common experimental region is preferable for the fairest comparison.

The MSTAR tests all demonstrated considerable savings in image processing, when compared to the number of images which must be collected and processed for ranking and selection or confidence interval construction. Note that sample sizes required to complete the Wald test can vary considerably; the random sample order for any given test may also lead to unusually large or small required sample sizes. However, the Wald test can significantly reduce required sample sizes to detect system differences most of the time. These reduced sample sizes translate into lower CPU, engineering, and data-collection costs, which can equate to hundreds of thousands of dollars. One would think that choosing to use the Wald test involves calculated risk, since the random data ordering could lead to unusually large sample size requirements; but since the experimenter can conduct the Wald test while collecting ranking and selection data and using the ranking and selection sample size as a stopping condition, he or she has nothing to lose in implementing the Wald test, provided the data is conducive to sequential testing. Data conducive to sequential testing is generally collected point by point or in small batches, though data points from large batches can also be randomly ordered and processed with the Wald test. In the case of large batches, ranking and selection or confidence intervals are usually faster to calculate, however, if sufficient data has been collected to accept or reject the hypothesis.

When implementing the Wald test, one should choose the parameters with special care. The analysis in Chapter 4 demonstrated that setting $\alpha = \beta$ and $u_0 = 1/u_1$ provides an fair comparison, i.e., the probability of the test selecting one system over the other is 0.5 when the system p_{ID}'s are equal. If an engineer wanted to test a new ATR system against one already installed in an intelligence center, for example, $\alpha < \beta$ and u-values which are not reciprocals could be appropriate parameter choices, but in that case, parameters should be chosen carefully. Choosing u-values closer to 1 allows for better resolution between the systems under examination, but also leads to higher sample size requirements, as the boundaries of the test move further apart. Estimates of the p_{ID}'s, the probabilities of success for the systems, also help in choosing u-values, as inserting the p_{ID} estimates into Equation (3.5) and then reinserting them in reverse order yields two starting u-values.

Overall, we obtained excellent results in implementing the Wald and Wald MSRB sequential testing procedures to compare two or four MSTAR configurations, assuming that the $\hat{p}_{ID}$'s for the all-points data reflected the true p_{ID}'s for each system. In the comparison of regions and peaks systems with the same FE and M modules (the P/R test), the $\hat{p}_{ID}$'s provided enough evidence to declare the regions configuration superior with 95% confidence; we exhibited nearly perfect accuracy at determining this difference with the Wald test, while reducing the necessary sample size considerably. The four-system comparisons consistently distinguished between the two best systems and two worst systems in each test, in accordance with the specified 0.03 sensitivity.

In addition to choosing the best system, WL wanted to ensure that at least one system meets the specified $\hat{p}_{ID} \geq 0.7$. Confidence intervals showed that systems R and P met this specified $\hat{p}_{ID}$, though none of the p_{ID}'s for systems A - H were statistically significantly ≥ 0.7. Since configuration R is a refined version of configuration G, one of the superior regions configurations, WL can safely choose to purchase configuration R and meet their performance specification of $\hat{p}_{ID} \geq 0.7$ as well as their goal of choosing the optimal configuration. (This assumes that a refined version of configuration E will have a p_{ID} within 0.03 or lower than p_R, an assumption which WL should verify before purchasing configuration R.)

The MSTAR test results have illustrated the success with which we can implement the Wald test to fairly compare any ATR systems, provided we clearly define the necessary degree of resolution (i.e., the sensitivity $|p_1 - p_2|$ which we want to detect) and the desired significance level, and balance these with the amount of available data. We can also simultaneously compare more than two systems by implementing the modified sequentially rejected Bonferroni approach. Therefore, the Wald test can provide an "honest broker" comparison of ATR systems, and allow system developers to choose the true superior ATR system which will maximize the safety of ground troops and other units which use ATR for safety and targeting.

6.2 Recommendations for Further Development

The methodology and procedure for the fair comparison of two or four systems can easily be extended to compare any number of systems. The Wald test could also be further explored to develop a procedure of unfair comparison of systems, such that system performance is weighted by cost, flexibility, or other quantitative measures and measures which can modeled in a quantitative form. This may be accomplished by setting the likelihood function $L(u)$ to a value other than 0.5, and using Wald's sample size formulas for the general case. For the fair comparison, we used the equations for our special case.

We treated the targets in the MSTAR data as equally likely, since the MSTAR developers want to build a mission-nonspecific system. In some experiments, we may want the best ATR system for a specific mission. For a mission-specific selection, we can simply modify the target sampling to favor the more likely targets in the mission scenario. Rather than selecting targets at random from the T, A, D, space, we can sample certain targets more frequently than others, depending on the probability of acquiring each target for the mission in question. The systems will thus be compared on their performance for targets which are representative of the mission, rather than assuming all targets equally likely.

Once an optimal ATR system has been chosen by implementing the sequential testing procedures developed in Chapter 4, one may want to model system performance over the T, A, D space, and use this model to improve the weak points of the system. This performance modeling could involve binary experimental design or neural networks, and presents a challenging problem in both the analysis and model verification and validation. One possible procedure for comparing systems while modeling performance is to design an experiment for modeling system performance, and then to randomly sample the design points for a Wald test, and continue sampling even after the Wald test has concluded until all replicates of all design points have been collected. This provides the engineer with information about the best system, as well designed experiment results for performance modeling analysis.

Finally, a potential research area lies in the construction of a three-way or more test to replace multiple pairwise comparisons. The error types would have to be expanded to include the incorrect rejection or acceptance of the third system, thus introducing a new parameter γ to complement α and β. For sequential analysis, the Wald test could be rederived to produce a three-dimensional graphical test with planes as boundaries. Proving the validity of such a test for both sequential analysis and conventional hypothesis testing presents an interesting challenge to the curious researcher, particularly if a three-way test can be found to be more powerful than multiple pairwise comparisons for three systems.

References

[1] Axtell, Mark. “MSTAR Data.” Veda, Inc., March 1996.

[2] Burns, Thomas J. Conversation about MSTAR, ATR, and “Honest Broker.” Wright Laboratory, Wright-Patterson AFB OH, 12 August 1996.

[3] ---. Slides, MSTAR Program Overview. Wright Laboratory, Wright-Patterson AFB OH, February 1996.

[4] Collett, D. *Modelling Binary Data.* London: Chapman & Hall, 1991.

[5] Dilsavor, Ronald. Conversation at AFIT concerning MSTAR data availability and independence. 21 May 1996.

[6] ---. “Wright Laboratory ADEPT Test Plan for MSTAR.” Phoenix AZ, March 1996.

[7] Gibbons, Jean Dickinson, Ingram Olkin, and Milton Sobel. *Selecting and Ordering Populations: A New Statistical Methodology.* New York: John Wiley and Sons, 1977.

[8] Hogg, Robert V. and Johannes Ledolter. *Applied Statistics for Engineers and Physical Scientists.* 2d ed. New York: Macmillan, 1992.

[9] Holm, Sture. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandanavian Journal of Statistics*, 6: 65-70 (1979).

[10] IEEE Standard Radar Definitions, IEEE Std 686-1990.

[11] Lockheed Martin. *MSTAR Design Documentation.* Denver CO, March 1996.

[12] Mason, Robert L., Richard F. Gunst, James L. Hess. *Statistical Design and Analysis of Experiments.* New York: John Wiley and Sons, 1989.

[13] Meer, David E. "Synthetic Aperture Radar." Course handouts, two-day SAR course. Air Force Institute of Technology, Wright-Patterson AFB OH, 1987-1992.

[14] Mendenhall, William, Dennis D. Wackerly, Richard L. Scheaffer. *Mathematical Statistics with Applications.* 4th ed. Belmont: Duxbury Press, 1990.

[15] Montgomery, D. C. *Design and Analysis of Experiments.* 3rd ed. New York: John Wiley and Sons, 1991.

[16] ---. Telephone conversations about experimental design. March-April 1996.

[17] Neter, John et al. *Applied Linear Statistical Models.* 4th ed. Chicago: Richard D. Erwin, Inc., 1989.

[18] Shaffer, Juliet Popper. "Modified Sequentially Rejective Multiple Test Procedures." *Journal of the American Statistical Association,* 81: 826-831 (1986).

[19] Wald, Abraham. *Sequential Analysis.* New York: John Wiley and Sons, 1947.

Appendix A

Confidence Interval Calculation when p_{ID} for Different Scenarios is Non-constant

In ATR, p_{ID} frequently varies between target types, and as aspect and depression angles change for each target. We may calculate a valid confidence interval which accounts for this as follows, for target t_i drawn from m possible targets (denoted $t_i \in \{t_1,\ldots,t_m\}$), aspect angle $a_j \in \{0^\circ,\ldots,360^\circ\}$, and depression angle $d_k \in \{10^\circ,\ldots,45^\circ\}$. We require data from N_{ijk} replicate observations for each scenario where N_{ijk} is as large as feasible, data availability permitting. Larger N_{ijk} values reduce interval width.

- Let s^{ijk} be the processing result of any ROI of a scenario (T = t_i, A = a_j, D = d_k) such that $s^{ijk} = 1$ signifies a correct target identification.
- Let $P(s^{ijk} = 1) = P^{ijk}$ for scenario (T = t_i, A = a_j, D = d_k)
- $\hat{P}^{ijk} = \frac{1}{N_{ijk}} \sum_{l=1}^{N_{ijk}} s_l^{ijk}$
- $E(\hat{P}^{ijk}) = P^{ijk}$ and $\mathrm{var}(\hat{P}^{ijk}) = \frac{P^{ijk}(1-P^{ijk})}{N_{ijk}}$
- For any target draw *ijk,* let P_{ijk} = P(T = t_i, A = a_j, D = d_k); in other words, P_{ijk} is the probability of choosing scenario (T = t_i, A = a_j, D = d_k). Note that probabilities of obtaining certain targets can change with different battle conditions, so to estimate an ATR system's performance for a certain battle, the analyst should properly adjust the P_{ijk}'s. For MSTAR, we assume that all scenarios are equally likely to be observed.

- Let $p = \mathrm{P}(s = 1) = \sum_{i}\sum_{j}\sum_{k} P^{ijk} P_{ijk}$

- Let $\hat{p} = \sum_{i}\sum_{j}\sum_{k} \hat{P}^{ijk} P_{ijk}$; then $\mathrm{E}(\hat{p}) = p$ and $\mathrm{var}(\hat{p}) = \sum_{i}\sum_{j}\sum_{k} P^{ijk}\left(1 - P^{ijk}\right)\frac{P_{ijk}^{2}}{N_{ijk}}$

- Resulting confidence interval for large *n*:

$$\sum_{i}\sum_{j}\sum_{k} \hat{P}^{ijk} P_{ijk} \pm z_{1-\alpha/2}\sqrt{\sum_{i}\sum_{j}\sum_{k} \hat{P}^{ijk}\left(1 - \hat{P}^{ijk}\right)\frac{P_{ijk}^{2}}{N_{ijk}}}$$

Appendix B

The Wald Sequential Probability Ratio Test

Pairwise sequential testing can be very helpful in identifying the best of a set of processes with minimal testing. However, in many cases, an experiment aims to test whether a single process meets a specification. A sequential test, the Wald sequential probability ratio test (SPRT), exists for this purpose.

The SPRT behaves similarly to the p_1 - p_2 test, except that the SPRT tests the simple hypothesis $H_0 \colon p = p_0$ against alternate $H_1 \colon p = p_1$. Also, we now count successes instead of t_2 and total samples m instead of $t_1 + t_2$. The critical values are calculated as follows [19:91]:

$$\text{lower bound:} \quad \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0}} + m \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0}} \qquad \text{(B.1)}$$

$$\text{upper bound:} \quad \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0}} + m \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0}} \qquad \text{(B.2)}$$

In the MSTAR case, we would test the compound hypothesis $H_0 \colon p < p_{spec}$ versus $H_1 \colon p \geq p_{spec}$, so choices of p_0 and p_1 slightly above and below p_{spec} respectively require engineering judgment. As before in the case of u, we need to select p_0 and p_1 such that we are indifferent in our decision when p is between p_0 and p_1.

The graphical implementation of the SPRT looks very similar to the other Wald test, and is performed in the same way.

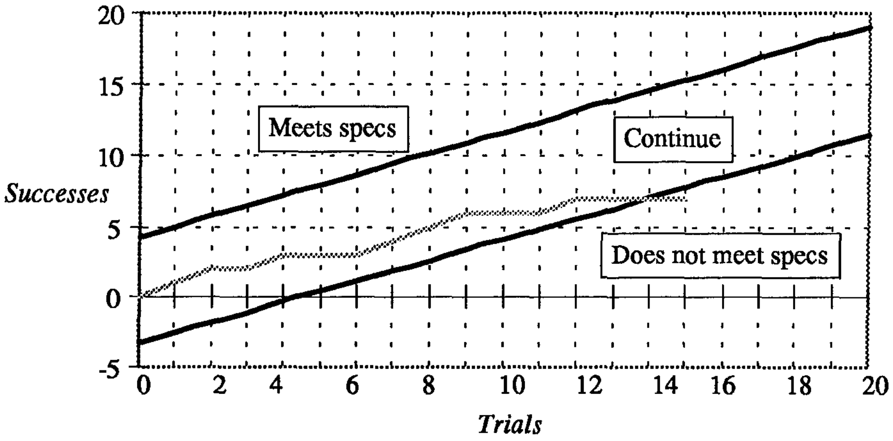

Figure B.1. Sample Wald SPRT Graph

To supplement the pairwise comparison testing plan, we could run a SPRT on each of the four processes while simultaneously performing the pairwise tests. This allows for the possibility that one or more of the processes will fail the specification tests before losing the pairwise tests, thus eliminating itself early from its pairwise tests and cutting testing requirements. However, preliminary experimentation showed that for the level of sensitivity required to determine whether a system meets specs, the comparison tests terminate well before the SPRT, thus making the redundant SPRT unnecessary. We therefore focused on the comparison test for selection of the superior system.

Appendix C

Slope of the Wald Test Boundaries

The following demonstration shows that the slope of the Wald test boundaries is equal to 0.5 whenever $u_0 = 1/u_1$, and equivalently, $u_1 = 1/u_0$.

$$\text{Slope } s = \frac{\log \dfrac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} \qquad \text{(C.1)}$$

Let $u_0 = 1/u_1$:

$$s = \frac{\log \dfrac{1+u_1}{1+\dfrac{1}{u_1}}}{\log u_1 - \log \dfrac{1}{u_1}} \qquad \text{(C.2)}$$

$$s = \frac{\log(1+u_1) - \log \dfrac{1+u_1}{u_1}}{\log u_1 - \log u_1^{-1}}$$

$$s = \frac{\log(1+u_1) - (\log(1+u_1) - \log u_1)}{\log u_1 - (-\log u_1)}$$

$$s = \frac{\log u_1}{2\log u_1} = \frac{1}{2}$$

The same arithmetic may be performed by substituting $u_1 = 1/u_0$ instead, with the same result.

Appendix D

Choosing u-values when $\alpha \neq \beta$

Though the best way of constructing the Wald test for fair comparison of two systems is to set $\alpha = \beta$ and $u_0 = 1/u_1$, a fair test can also be constructed when α and β are different. To choose suitable *u*-values, we can perform an iterative simulation procedure similar to the one used for the sensitivity experiments. This case reflects WL's specification of $\beta = 0.05$ and $\beta = 0.10$, and is a summary of the preliminary experiments performed before we concluded that we should set $\alpha = \beta$ and $u_0 = 1/u_1$.

The simulation was conducted with the following procedure.

1. Choose the first (p_1,p_2) pair shown in Table 4.1, (0.75,0.72). Let $(\alpha, \beta) = (0.05,0.10)$ per WL specification. Let $(u_0, u_1) = (0.86,1.16)$ as a starting setting.
2. Randomly generate 300 Bernoulli trials for $p_1 = 0.75$ and 300 for $p_2 = 0.72$, using the `binornd` command in the MATLAB® Statistics Toolbox. (WL chose to process a maximum of 300 images per configuration.)
3. Run the Wald test on the simulated data using a MATLAB® program.
4. Record whether the test correctly chose the system 1 (or the system with $p = 0.75$) as superior.
5. Starting at Step 1, repeat the simulation 99 more times, and average the 100 results to obtain a percentage of runs for which the test correctly chose the best system. Also record percentage of incorrect and inconclusive tests.
6. Repeat Steps 1-5 with all other (p_1, p_2) pairs shown in Table 4.1.
7. Reverse the system order and repeat Steps 1-6, such that the first (p_1, p_2) pair is (0.72, 0.75). This ensures that correct testing does not depend on system order. Average these results with those obtained in Step 5 for each (p_1, p_2) pair.

8. Repeat Steps 2-7 for a range of u_0 and u_1 values.
9. Compare test performance for each u-value combination, and select the (u_0, u_1) pair with the best test performance in the experimental region.

In initial tests with u_0 and u_1 set at 0.86 and 1.16 respectively, more than half of the tests were inconclusive with only 300 samples. Trial and error showed increases in test performance when both parameters were increased. Based on this preliminary information, we chose a range of nine u_0 and six u_1 values and tested the system at all value combinations, such that $0.82 < u_0 < 0.96$ and $1.5 < u_1 < 1.8$. The surface which resulted from the range simulations follows in Figure D.1.

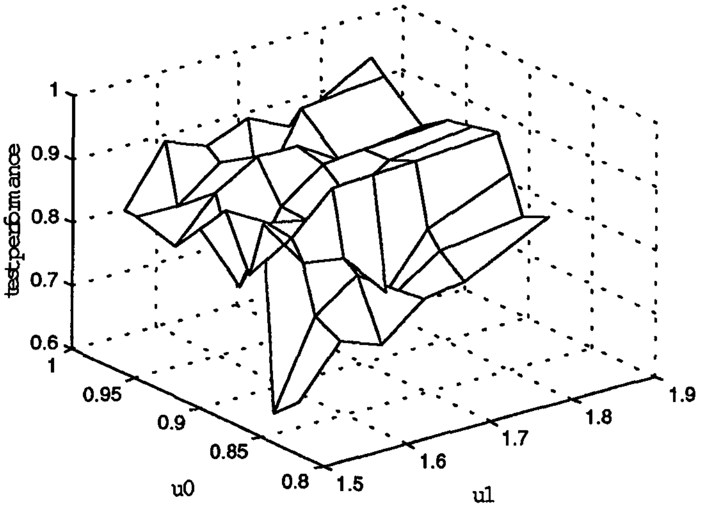

Figure D.1. Wald test performance for varying parameter levels

This surface is difficult to read well. To produce a smoother plot, we input the performance data into a neural network and used simulated output to create a smoothed surface, which appears imprecise but is useful for our needs.

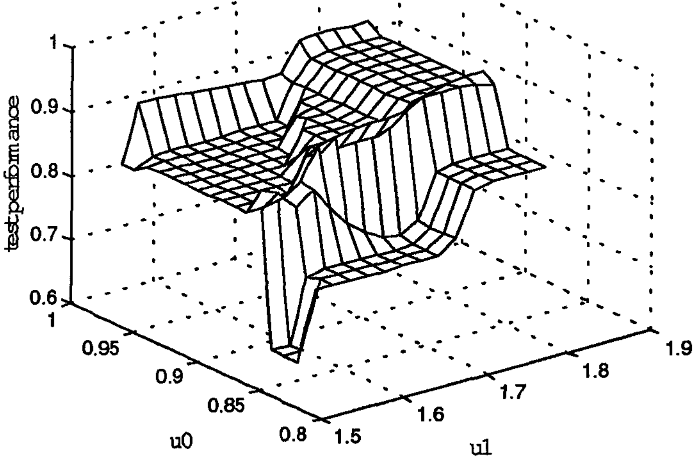

Figure D.2. Wald test performance for varying parameter levels (smoothed)

Performance clearly increases as u_1 increases. Interaction between the parameters also seems to exist. We will move in the direction of increasing u_1 and investigate further to choose the preferred value.

We move to the right side of the surface and simulate on a new region, including six u_0 and eight u_1 values such that $0.86 < u_0 < 0.98$ and $1.5 < u_1 < 3$. The following surface resulted:

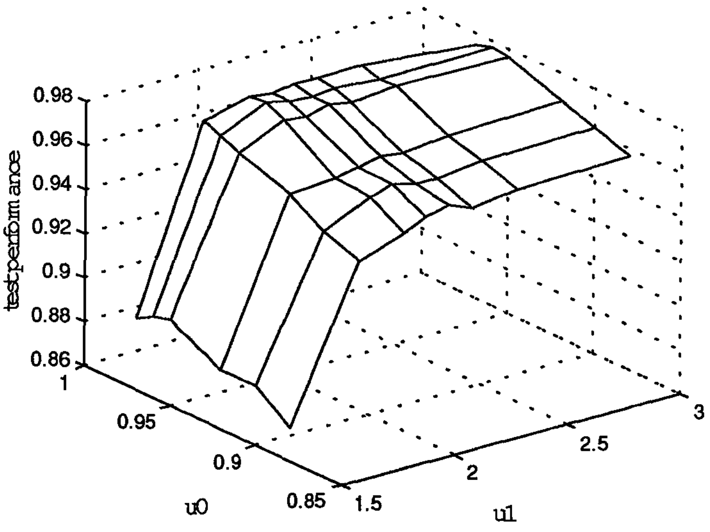

Figure D.3. Wald test performance for various parameter values

Here, we see that setting (u_0, u_1) to (0.96,2) achieves excellent performance, with an average of 98% of Wald tests successfully choosing the known superior configuration, and very few tests producing inconclusive or incorrect results. We should bear in mind that the random data generated may actually have had sample p_{ID}'s which varied from the true p_{ID}'s, yet the tests still mostly produced correct results as for the true p_{ID}'s.

The surface plot shows that performance may increase for u_1 values over 2. However, we should revert back to Wald's directions for parameter selection and note that by setting a parameter at 2, we already imply that we are testing that one procedure is twice as "efficient" as another. The favoring of the status quo in the hypothesis structure previously discussed makes "efficient" a relative term, mostly in the sense of cost effectiveness of the new process installation and operational savings versus the old process operational costs. This conditional nature of the term "efficiency" means that we are not restricted to keeping u_1 under 2, but in the interest of staying somewhat near Wald's guidelines, we can forsake very slight improvements in test performance which we could gain with higher levels of u_1. Increased u_1 levels simply suffer from diminished returns in terms of test performance.

Appendix E

u-values for Test Sensitivity Study

0.01		**0.03**		**0.04**		**0.05**	
p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
0.66	0.67	0.65	0.68	0.65	0.69	0.64	0.69
0.67	0.68	0.66	0.69	0.66	0.70	0.65	0.70
0.68	0.69	0.67	0.70	0.67	0.71	0.66	0.71
0.69	0.70	0.68	0.71	0.68	0.72	0.67	0.72
0.70	0.71	0.69	0.72	0.69	0.73	0.68	0.73
0.71	0.72	0.70	0.73	0.70	0.74	0.69	0.74
0.72	0.73	0.71	0.74	0.71	0.75	0.70	0.75
0.73	0.74	0.72	0.75	0.72	0.76	0.71	0.76

0.06		**0.07**		**0.08**		**0.10**	
p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
0.63	0.69	0.63	0.70	0.62	0.70	0.62	0.72
0.64	0.70	0.64	0.71	0.63	0.71	0.63	0.73
0.65	0.71	0.65	0.72	0.64	0.72	0.64	0.74
0.66	0.72	0.66	0.73	0.65	0.73	0.65	0.75
0.67	0.73	0.67	0.74	0.66	0.74	0.66	0.76
0.68	0.74	0.68	0.75	0.67	0.75	0.67	0.77
0.69	0.75	0.69	0.76	0.68	0.76	0.68	0.78
0.70	0.76	0.70	0.77	0.69	0.77	0.69	0.79

Appendix F

Sample Sizes for Confidence Intervals when α = 0.05

True		0.01	*True*		0.02	*True*		0.03	*True*		0.04
p1	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*
0.66	0.67	17114	0.65	0.67	4308	0.65	0.68	1900	0.65	0.69	1060
0.66	0.67	17114	0.66	0.68	4245	0.66	0.69	1871	0.66	0.70	1043
0.67	0.68	16852	0.67	0.69	4178	0.67	0.70	1840	0.67	0.71	1025
0.68	0.69	16576	0.68	0.70	4107	0.68	0.71	1808	0.68	0.72	1006
0.69	0.70	16284	0.69	0.71	4032	0.69	0.72	1773	0.69	0.73	987
0.70	0.71	15977	0.70	0.72	3953	0.70	0.73	1738	0.70	0.74	966
0.71	0.72	15654	0.71	0.73	3870	0.71	0.74	1700	0.71	0.75	945
0.72	0.73	15316	0.72	0.74	3784	0.72	0.75	1661	0.72	0.76	922
Average		16361	*Average*		4059	*Average*		1786	*Average*		994

True		0.05	*True*		0.06	*True*		0.07	*True*		0.08
p1	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*
0.64	0.69	683	0.64	0.70	470	0.63	0.70	347	0.63	0.71	263
0.65	0.70	672	0.65	0.71	462	0.64	0.71	342	0.64	0.72	259
0.66	0.71	661	0.66	0.72	455	0.65	0.72	336	0.65	0.73	255
0.67	0.72	650	0.67	0.73	446	0.66	0.73	330	0.66	0.74	250
0.68	0.73	637	0.68	0.74	437	0.67	0.74	324	0.67	0.75	245
0.69	0.74	624	0.69	0.75	428	0.68	0.75	318	0.68	0.76	240
0.70	0.75	611	0.70	0.76	419	0.69	0.76	311	0.69	0.77	235
0.71	0.76	597	0.71	0.77	409	0.70	0.77	303	0.70	0.78	229
Average		642	*Average*		441	*Average*		327	*Average*		247

True		0.09	*True*		0.10	*True*		0.11	*True*		0.12
p1	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*	*p1*	*p2*	*Size*
0.62	0.71	209	0.62	0.72	168	0.61	0.72	140	0.61	0.73	116
0.63	0.72	206	0.63	0.73	165	0.62	0.73	137	0.62	0.74	114
0.64	0.73	203	0.64	0.74	162	0.63	0.74	135	0.63	0.75	112
0.65	0.74	199	0.65	0.75	159	0.64	0.75	133	0.64	0.76	110
0.66	0.75	195	0.66	0.76	156	0.65	0.76	130	0.65	0.77	108
0.67	0.76	191	0.67	0.77	153	0.66	0.77	127	0.66	0.78	106
0.68	0.77	187	0.68	0.78	150	0.67	0.78	125	0.67	0.79	103
0.69	0.78	183	0.69	0.79	146	0.68	0.79	122	0.68	0.80	101
Average		197	*Average*		157	*Average*		131	*Average*		109

Appendix G

Wald Test Parameter Calculation

When $\alpha = \beta$ and $u_0 = 1/u_1$, we can calculate the exact tradeoff in α (and therefore β) and sensitivity for a fixed expected sample size by holding the slopes and intercepts of the test boundaries constant. The following derivations demonstrate this.

The Wald test for u_0 - u_1 has four variables and three equations, after decomposing the boundary equations:

- variables: α, β, u_0, u_1
- equations: upper bound intercept, lower bound intercept, slope

Given a variable vector (α, β, u_0, u_1), suppose that we want to the u_0 and u_1 values for a given new (α, β) pair, as in the MSRB Wald test, such that we maintain the same slope and intercepts as in the original case. We will refer to the original vector (α, β, u_0, u_1) and the new vector as (α', β', u_0', u_1'). The values α' and β' are specified (for example, in the four-system comparison test, (α', β') = (α/6, β/6)).

We start by setting the original equations equal to the new equations:

$$\text{Slope:} \qquad s = \frac{\log \frac{1+u_1}{1+u_0}}{\log u_1 - \log u_0} = \frac{\log \frac{1+u_1'}{1+u_0'}}{\log u_1' - \log u_0'}$$

$$\text{Upper intercept:} \qquad \mathrm{U} = \frac{\log \frac{1-\beta}{\alpha}}{\log u_1 - \log u_0} = \frac{\log \frac{1-\beta'}{\alpha'}}{\log u_1' - \log u_0'}$$

$$\text{Lower intercept:}\qquad \mathrm{L} = \frac{\log\dfrac{\beta}{1-\alpha}}{\log u_1 - \log u_0} = \frac{\log\dfrac{\beta'}{1-\alpha'}}{\log u_1' - \log u_0'}$$

Since the left-hand sides are all known when the original vector (α, β, u_0, u_1) is given, we will rename them S, U, and L, respectively (slope, upper intercept, lower intercept.)

Beginning with the slope and upper intercept equations, we can solve each in terms of the denominator $\log u_1' - \log u_0'$:

$$\text{Slope:}\qquad \log u_1' - \log u_0' = \frac{\log\dfrac{1+u_1'}{1+u_0'}}{S}$$

$$\text{Upper intercept:}\qquad \log u_1' - \log u_0' = \frac{\log\dfrac{1-\beta'}{\alpha'}}{U}$$

It then follows that

$$\frac{\log\dfrac{1-\beta'}{\alpha'}}{U} = \frac{\log\dfrac{1+u_1'}{1+u_0'}}{S}$$

Since α' and β' are also known, solve this equation for u_1 as a function of u_0 as follows:

$$\frac{S}{U}\log\frac{1-\beta'}{\alpha'} = \log\frac{1+u_1'}{1+u_0'}$$

$$\log\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} = \log\frac{1+u_1'}{1+u_0'}$$

$$\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} = \frac{1+u_1'}{1+u_0'}$$

$$\left(1+u_0'\right)\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} - 1 = u_1'$$

$$u_1' = u_0'\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} + \left(\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} - 1\right) \tag{G.1}$$

We have now solved for u_1' in terms of u_0', such that the linear Equation (G.1) provides a relationship to find u-value pairs which give the same slope and upper intercept as for the original vector. However, notice that if we perform the exact same procedure using the lower intercept equation, we obtain Equation (G.2):

$$u_1' = u_0'\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}} + \left(\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}} - 1\right) \tag{G.2}$$

Setting the u_1' values equal:

$$u_0'\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} + \left(\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} - 1\right) = u_0'\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}} + \left(\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}} - 1\right)$$

$$u_0'\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} + \left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} = u_0'\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}} + \left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}}$$

$$(1+u_0')\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{S}{U}} = (1+u_0')\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{S}{L}}$$

$$\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{1}{U}}=\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{1}{L}}$$

Using (α', β') = (α/6, β/6) = (0.05/6, 0.1/6) and (L, U) = (-2.6675, 3.4247) from the 4-system and 2-system comparison respectively with (u_0,u_1) = (0.86,2):

$$\left(\frac{0.9833}{0.0083}\right)^{\frac{1}{3.4247}}=\left(\frac{0.0167}{0.9917}\right)^{\frac{1}{-2.6675}}$$

$$4.0270 \neq 4.6263$$

Therefore, by contradiction, we cannot necessarily solve for a new vector (α', β', u_0', u_1') when given an initial vector (α, β, u_0, u_1) and new parameters (α', β'), unless $\alpha=\beta$. In this case, the lower intercept is the negative of the upper intercept, since when $\alpha=\beta$, the following holds:

$$\log\frac{1-\beta}{\alpha}=\log\frac{1-\alpha}{\alpha}=-\log\frac{\alpha}{1-\alpha}=-\log\frac{\beta}{1-\alpha}$$

The equations for the upper and lower intercepts, respectively, are

$$\left(\frac{\log\frac{1-\beta}{\alpha}}{\log u_1-\log u_0},\frac{\log\frac{\beta}{1-\alpha}}{\log u_1-\log u_0}\right)$$

Therefore, since $\log\frac{1-\beta}{\alpha}=-\log\frac{\beta}{1-\alpha}$ when $\alpha=\beta$, U = -L.

Given that U = -L and $\alpha=\beta$, so also $\alpha'=\beta'$:

$$\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{1}{U}}=\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{1}{L}}$$

$$\left(\frac{1-\beta'}{\alpha'}\right)^{\frac{1}{U}}=\left(\frac{1-\alpha'}{\alpha'}\right)^{\frac{1}{U}}=\left(\frac{\alpha'}{1-\alpha'}\right)^{\frac{1}{-U}}=\left(\frac{\alpha'}{1-\alpha'}\right)^{\frac{1}{L}}=\left(\frac{\beta'}{1-\alpha'}\right)^{\frac{1}{L}}$$

Therefore, if $\alpha = \beta$, we can obtain the same intercepts with (α', β'). We can also obtain the same slope if $u_0 = 1/u_1$, which always provides a slope of 0.5, as shown in Appendix C.

Appendix H

MATLAB® code for Peaks Wald Test

This code can easily be adapted for the Regions and P/R test by modifying the names of the loaded data files and the data names.

```
close all;
clear all;
clc;

% initialize variables

u0vals = ones(1,6)*.867;
u1vals = 1./u0vals;
alpha = 0.05;
beta = alpha;
dropped = 0;
cnflag=0;

% SET THESE THREE PARAMETERS
nruns=2; % number of test runs - use 1 to do only one
nruns = input('Number of runs [1]: ');
if isempty(nruns),nruns = 1; end;

% DATA CHOICES
simflag = 1;  % 1 for simulate, 0 for randomize
simflag = input('Enter 1 to simulate or 0 to randomize data [1]: ');
if isempty(nruns),simflag = 1; end;

disp('Choose data set: ');
disp('1 - regions/peaks, all points');
disp('2 - regions/peaks, common sets');
disp('3 - peaks, common sets');
disp('4 - peaks, all points');
disp('5 - regions, common sets');
disp('6 - regions, all points');
disp('7 - special Pids for simulation, coded into program');
dataflag = input('Enter data choice [7]: ');
if isempty(dataflag), dataflag = 7; end;
%dataflag = 7;       % 1 for XR,XP,
                     %2 for regions,peaks
                     % 3 for PA - PD
                     % 4 for XA - XD
                     % 5 for RE - RH
                     % 6 for XE - XH
                     % 7 for testing

% load data - all-points sets

if dataflag == 1, load XR.txt; load XP.txt; lvec = [length(XR) length(XP)]; end;
if dataflag == 4, load XA.txt; load XB.txt; load XC.txt; load XD.txt; lvec = [length(XA) length(XB) length(XC)
length(XD)]; end;
if dataflag == 6, load XE.txt; load XF.txt; load XG.txt; load XH.txt; lvec = [length(XE) length(XF) length(XG)
length(XH)]; end;
if dataflag == 3, load PA.txt; XA = PA(:,7); load PB.txt; XB = PB(:,7); load PC.txt; XC = PC(:,7); load PD.txt; XD =
PD(:,7); lvec = length(PA); end;
if dataflag == 5, load RE.txt; XE = RE(:,7); load RF.txt; XF = RF(:,7); load RG.txt; XG = RG(:,7); load RH.txt; XH =
RH(:,7); lvec = length(RE); end;
if dataflag == 2, load regions.txt; XR = regions(:,7); load peaks1.txt; XP = peaks1(:,7); lvec = length(XP); end;
if dataflag == 7, lvec = 1; end;
datalength = min(lvec);

if (dataflag == 1|dataflag == 2), pids = [mean(XR) mean(XP)]; names = ['R' 'P']; end;
if (dataflag == 3|dataflag == 4), pids = [mean(XA) mean(XB) mean(XC) mean(XD)]; names = ['A' 'B' 'C' 'D']; end;
if (dataflag == 5|dataflag == 6), pids = [mean(XE) mean(XF) mean(XG) mean(XH)]; names = ['E' 'F' 'G' 'H']; end;
if dataflag == 7, pids = [.715 .685 .685 .685]; names = ['J' 'K' 'L' 'M']; end;


% BEGIN RUNS

for runs=1:nruns
clear dropped dropme cnflag cnumc i j;
cnflag=0;
dropped=0;

% randomize data order or simulate data

% for lock-step
if dataflag == 2, rawdata = [XR XP]; end;
if dataflag == 3, rawdata = [XA XB XC XD]; end;
if dataflag == 5, rawdata = [XE XF XG XH]; end;

if simflag == 0
        r = rand(4,datalength);
        [r1,index] = sort(r(1,:));
        [r2,index2] = sort(r(2,:));
        [r3,index3] = sort(r(3,:));
        [r4,index4] = sort(r(4,:));

        for i=1:length(index)
%                tdata(i,:) = rawdata(index(i),:); % for lock-step
                 tdata(i,:) = [XR(index(i)) XP(index2(i))]; % for random draw
        end;
else
        Npts = 10000; % sample size to generate
        for i=1:length(pids)
                tdata(:,i) = binornd(1,pids(i),Npts,1);
        end;
end;

% randomize config order

rr = rand(1,size(tdata,2));
[nt,v] = sort(rr);
for i=1:length(v)
        rdata(:,i) = tdata(:,v(i));
end;
for kk=1:length(v), name(kk)=names(v(kk)); end;
```

```
pt=mean(rdata);

numsamps = size(rdata,1);  % number of samples for each configuration
numconf = length(nt); % number of configurations
nconf = length(nt); % number of configs - won't be decremented
pairs = combnk(inx(numconf),2);
pairix = inx(size(pairs,1));
cnumc = length(pairix);  % number of comparisons

for i=1:(numconf)
        numc(numconf-i+1) = size(combnk(inx(i),2),1);  % number of comparisons after each drop
end;

% Construct Wald intercepts, slopes on first run

if runs==1
        denom = log(u1vals) - log(u0vals);
        slope = log((1 + u1vals)./(1 + u0vals))./denom;
        for i = 1:numsamps
        for j = 1:size(pairs,1)
                fline(i,j) = log((beta/j) / (1 - (alpha/j)))/denom(j) + i*slope(j);
                pline(i,j) = log((1 - (beta/j)) / (alpha/j))/denom(j) + i*slope(j);
        end;
        end;
end; % if runs==1

% plot pass/fail lines
%plot(linspace(1,30,30),fline(1:30,:));
%hold;
%plot(linspace(1,30,30),pline(1:30,:));
%title('Wald Bounds for Decreasing Alpha Values','fontname','timesnewroman');
%axis([1 30 0 25]);
%xlabel('t1 + t2','fontname','timesnewroman');
%ylabel('t2','fontname','timesnewroman');
%disp('Press any key');
%pause;
%close;

% The Wald Sequential Test for p1 - p2

clear t2 t;

t2(1,:) = zeros(1,numc);
t = t2;
report = t2;
rc = ones(1,numc);  % row counter for pline/fline

for j=1:numsamps
rcflag = zeros(1,length(pairix));
for i=1:max(pairix)
        % check for nonmatched data pairs
        if pairix(i)>0
                if rdata(j,pairs(i,1)) > rdata(j,pairs(i,2))
                        t(i) = t(i) + 1;
                        t2(rc(i),i) = max(t2(:,i));
                        rcflag(i) = 1;
                elseif rdata(j,pairs(i,1)) < rdata(j,pairs(i,2))
```

```
                        t(i) = t(i) + 1;
                        t2(rc(i),i) = max(t2(:,i)) + 1;
                        rcflag(i) = 1;
                end;

        % compare to pass/fail lines, update report
        if rcflag(i)==1
                if t2(rc(i),i) > pline(rc(i),cnumc), report(i) = 2;
                elseif t2(rc(i),i) < fline(rc(i),cnumc), report(i) = 1; end;
                tpassline(rc(i),i) = pline(rc(i),cnumc);
                tfailline(rc(i),i) = fline(rc(i),cnumc);
                rc(i) = rc(i) + rcflag(i);
        end; % rcflag(i)==1

        end; % for pairix(i)>0

end; % for i=1:max(pairix) to get reports

% drop losing configs
for k=1:max(pairix)
        dropme=0;
        if report(k) == 1, dropme = pairs(k,2);
        elseif report(k) == 2, dropme = pairs(k,1); end;

        % check for no drop of 0 and if config was already dropped
        dflag=0;
        if dropme>0&numconf>1, dflag=1; end;
        for g=1:length(dropped)
                if dropme==dropped(g)|dropme==0, dflag=0; end;
        end;
        if dflag==1
                if dropped == 0, dropped = dropme;
                else dropped = [dropped dropme]; end;
                numconf = numconf - 1;
        end;

        % drop loser
        if numconf>1
        for m=1:size(pairs,1)
        if dflag==1
                if (pairs(m,1)==dropme)|(pairs(m,2)==dropme)
                        pairs(m,:) = [0 0];
                        pairix(m) = 0;
                        cnumc = cnumc - 1;
                end;
        end; % if dflag==1
        if cnumc == 1
                lastix = max(pairix);
                lastpairs = pairs(max(pairix),:);
        end; % if cnumc==1
        end; % for m=1:size(pairs,1)
        end; % for if numconf>1

if numconf==1
        cnflag=1;
        win = num2str(sum(inx(nconf))-sum(dropped));
        break;
end;
```

```
 end; % for k=1:max(pairix) for losing config
 if numconf==1, break, end;

 end; % for j=1:numsamps


clc;
disp(['RUN # ' num2str(runs)]);
if cnflag>0
        compl(runs) = 1;
        lastt2 = t2(:,lastix);
        for l=2:length(lastt2)
                if lastt2(l)<lastt2(l-1),lastt2=[lastt2(1:l-1)]; break; end;
        end;
        lastt = inx(length(lastt2));
        lpair1 = num2str(name(lastpairs(1)));
        lpair2 = num2str(name(lastpairs(2)));
        stp = length(lastt);
        x = linspace(1,stp,stp);
        plot(lastt,lastt2,'r',x,tpassline(1:stp,lastix),'g',x,tfailline(1:stp,lastix),'g');
        xlabel('t1 + t2','FontName','TimesNewRoman','FontSize',14);
        ylabel('t2','FontName','TimesNewRoman','FontSize',14);
        title(['Configuration ' lpair1 ' (bottom) vs. Configuration ' lpair2 '
(top)'],'fontsize',14,'FontName','TimesNewRoman');
        axis([1 stp 0 max(tpassline(1:stp,lastix))+2]);
        titlefnt;

        maxs = num2str(j);
        disp([maxs ' total samples needed to choose #' win ' as the best configuration.']);
        disp('Drop order: ');
        disp(dropped);
        [r,s]=sort(pt);
        disp('True order (randomized index):');
        disp(s);
        disp(r);
        disp('True pid for order data was loaded: ');
        disp(mean(tdata));
        disp('Key - index order over load order');
        disp(inx(length(v)));
        disp(['    'names(v(1)) '    'names(v(2))]); % '    'names(v(3)) '    'names(v(4))]);


elseif j==numsamps
        win=num2str(0);
        compl(runs) = 0;
        disp('More samples required');
        disp('Configurations dropped:');
        disp(dropped);
        [r,s]=sort(pt);
        disp('True order:');
        disp(s);
        disp(r);
        disp('True pid for order data was loaded: ');
        disp(mean(tdata));
        disp('Key - index order over load order');
        disp(inx(length(v)));
        disp(['    'names(v(1)) '    'names(v(2))]); % '    'names(v(3)) '    'names(v(4))]);
        disp('Number remaining:');
        disp(numconf);
        if cnumc==1
                lastt2 = t2(:,lastix);
                for l=2:length(lastt2)
                        if lastt2(l)<lastt2(l-1),lastt2=[lastt2(1:l-1)]; break; end;
                end;
                lastt = inx(length(lastt2));
                lpair1 = num2str(name(lastpairs(1)));
                lpair2 = num2str(name(lastpairs(2)));
                stp = length(lastt);
                x = linspace(1,stp,stp);
                plot(lastt,lastt2,'r',x,tpassline(1:stp,lastix),'b',x,tfailline(1:stp,lastix),'b');
                xlabel('t1 + t2','FontName','TimesNewRoman','FontSize',14);
                ylabel('t2','FontName','TimesNewRoman','FontSize',14);
                title(['Configuration ' lpair1 ' (bottom) vs. Configuration ' lpair2 '
(top)'],'fontsize',14,'FontName','TimesNewRoman');
                axis([1 stp 0 stp+2]);
        end; % if cnumc==1
end;

% get run reports

dtemp=[dropped zeros(1,nconf-length(dropped))];
repdrop(runs,:)=dtemp;
reptrue(runs,:)=s;
repkey(runs,:)=v;
repsamps(runs)=j;
repwin(runs,1)=(str2num(win)*(str2num(win)>0)) + 0;

end; % for runs=1:nruns

% collect sample size stats on complete runs only

for reep=1:length(repsamps)
        if compl(reep)>0,ssizs=[ssizs;repsamps(reep)]; end;
%       if repsamps(reep)~=length(rdata),ssizs=[ssizs;repsamps(reep)]; end;
end;

trepsamps = repsamps';
repcompl = mean(compl); % pct of time test completed with less than avail. samples
pctcorr = mean(compl'.*(repwin==reptrue(:,size(reptrue,2))));
pct2 = mean(compl'.*(repwin==reptrue(:,size(reptrue,2)-1)));
ssizavg = mean(ssizs);
ssizstdv = std(ssizs);
ssizcil = cil(.05,ssizavg,ssizstdv,nruns);
ssizciu = ciu(.05,ssizavg,ssizstdv,nruns);
ssizci=[ssizcil ssizciu]; % 95% ci of sample size
tpctcorr = pctcorr/repcompl; % pct of completed test which chose system with highest pid
tpctcorr2 = (pctcorr+pct2)/repcompl; pct of completed test which chose system with 1st or 2nd highest pid

save filename.mat pctcorr tpctcorr tpctcorr2 repcompl ssizs ssizci ssizavg ssizstdv;


end % program
```

Appendix I

MSTAR Data Summaries

I.1. 95% Confidence intervals for p_{ID} for all MSTAR systems

	All data points			*Common data points*		
System	Sample size	$\hat{p}_{ID}$	Interval	Sample size	$\hat{p}_{ID}$	Interval
A	240	0.6167	(0.5221, 0.6836)	141	0.6028	(0.5552, 0.6782)
B	194	0.6237	(0.5294, 0.6904)	141	0.6099	(0.5555, 0.6919)
C	237	0.6498	(0.5368, 0.6973)	141	0.6170	(0.5891, 0.7105)
D	226	0.6593	(0.6039, 0.7578)	141	0.6809	(0.5975, 0.7211)
E	283	0.7103	(0.6564, 0.7699)	244	0.7131	(0.6574, 0.7631)
F	281	0.6797	(0.6132, 0.7310)	244	0.6721	(0.6252, 0.7343)
G	259	0.7336	(0.6781, 0.7891)	244	0.7336	(0.6798, 0.7874)
H	272	0.6618	(0.5919, 0.7114)	244	0.6516	(0.6055, 0.7180)
R	1152	0.7708	(0.7465,0.7951)	962	0.7994	(0.7781,0.8206)
P	1048	0.7376	(0.7110,0.7642)	962	0.7297	(0.7062,0.7533)

I.2. Difference Intervals and Sample Sizes for Pairwise Comparisons of Systems

Interval calculation [14:360]:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{I.1}$$

Sample size required to declare $|p_1 - p_2|$ significant, where $n = n_1 = n_2$:

$$n = \left(z_{1-\alpha/2} \frac{\sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}}{(\hat{p}_1 - \hat{p}_2)} \right)^2 \tag{I.2}$$

All of the following intervals contain zero, and therefore demonstrate a statistically insignificant difference, except for the ones in bold.

All data points, a = 0.05

A v B	A v C	A v D
-0.0989 0.0848	-0.1196 0.0533	-0.1298 0.0446
B v C	B v D	C v D
-0.1174 0.0652	-0.1276 0.0565	-0.0961 0.0772
E v F	E v G	E v H
-0.0454 0.1065	-0.0988 0.0521	-0.0287 0.1257
F v G	F v H	G v H
-0.0060 0.0227	-0.0610 0.0963	-0.0060 0.1497
	R v P	
	-0.0028 0.0692	

All data points, a = 0.05/6

A v B	A v C	A v D
-0.1306 0.1165	-0.1495 0.0832	-0.1600 0.0748
B v C	B v D	C v D
-0.1490 0.0968	-0.1594 0.0883	-0.1261 0.1072
E v F	E v G	E v H
-0.0717 0.1328	-0.1249 0.0782	-0.0554 0.1524
F v G	F v H	G v H
-0.0329 0.0493	-0.0883 0.1234	-0.0329 0.1766
	R v P	
	-0.0153 0.0817	

Common data points, a = 0.05

A v B	A v C	A v D
-0.1211 0.1069	-0.1280 0.0996	-0.1896 0.0334
B v C	B v D	C v D
-0.1208 0.1066	-0.1824 0.0404	-0.1751 0.0473
E v F	E v G	E v H
-0.0408 0.1228	-0.0999 0.0589	-0.0209 0.1439
F v G	F v H	**G v H**
-0.1424 0.0194	-0.0634 0.1044	**0.0004 0.1636**
	R v P	
	0.1075 0.0319	

Common data points, a = 0.05/6

A v B	A v C	A v D
-0.1606 0.1464	-0.1674 0.1390	-0.2283 0.0721
B v C	B v D	C v D
-0.1601 0.1459	-0.2209 0.0789	-0.2135 0.0857
E v F	E v G	E v H
-0.0691 0.1511	-0.1273 0.0863	-0.0495 0.1725
F v G	F v H	G v H
-0.1704 0.0474	-0.0925 0.1335	-0.0278 0.1918
	R v P	
	0.0188 0.1206	

I.3. Minimum Sample Size Requirements for Difference Detection

For information only. In an experiment, we must precalculate the sample size and independently sample the data to use confidence intervals for analysis. These tables demonstrate the sample sizes that would be required to detect differences between the MSTAR configurations if the WL data p_{ID}'s are perfect estimates of the true p_{ID}'s.

All data points, a = 0.05

A v B	A v C	A v D
36,410	1,622	976
B v C	B v D	C v D
2,607	1,396	19,452
E v F	E v G	E v H
1,743	2,827	702
F v G	F v H	G v H
546	5,264	312
	R v P	
	1,290	

All data points, a = 0.05/6

A v B	A v C	A v D
65,972	2,939	1,768
B v C	B v D	C v D
4,723	2,530	35,245
E v F	E v G	E v H
3,158	5,122	1,271
F v G	F v H	G v H
990	9,539	565
	R v P	
	2,338	

Common data points, a = 0.05

A v B	A v C	A v D
36,376	9,063	288
B v C	B v D	C v D
36,138	347	427
E v F	E v G	E v H
971	3,657	438
F v G	F v H	G v H
422	4,090	241
	R v P	
	283	

Common data points, a = 0.05/6

A v B	A v C	A v D
65,911	16,422	521
B v C	B v D	C v D
65,480	629	773
E v F	E v G	E v H
1,760	6,625	794
F v G	F v H	G v H
765	7,410	437
	R v P	
	512	

Vita

Second Lieutenant Anne E. Catlin [redacted] She grew up in Damascus, Maryland, excluding the years 1986-1989. At that time she lived in the County of Cheshire, England, and attended Penrhos College in Colwyn Bay, Wales. After three years of mandatory knee socks and four-o'clock tea, she returned to Maryland and graduated from Damascus High School in 1991. She then attended Cornell University in Ithaca, New York for four years, during which time she tried to major in computer science and mechanical engineering before settling into operations research. She graduated in May 1995 with a B.S. in Operations Research and Industrial Engineering, with concentrations in Mechanical Engineering and Classics. She entered the Air Force Institute of Technology in August 1995, and after graduation in March 1997, reported to the 422 Test and Evaluation Squadron at Nellis AFB, Nevada. Among her hobbies, Lieutenant Catlin enjoys most outdoor activities, reading, Cornell hockey, and wine tasting, and is a die-hard Baltimore Orioles fan.

[redacted]

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
March 1997

3. REPORT TYPE AND DATES COVERED
Master's Thesis

4. TITLE AND SUBTITLE
SYSTEM COMPARISON PROCEDURES FOR AUTOMATIC
TARGET RECOGNITION SYSTEMS

5. FUNDING NUMBERS

6. AUTHOR(S)
Anne E. Catlin, Second Lieutenant, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
2950 P Street
WPAFB, OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GOR/ENS/97M-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
WL/AAC (AFMC)
Building 18F 5th Street
WPAFB, OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Estimating the performance of an automatic target recognition (ATR) system in terms of probability of successful target identification involves extensive image collection and processing, which can be very time-consuming and expensive. We investigate the Wald sequential test for the difference in two proportions as a sample size-reducing alternative to ranking and selection and the classical method of comparing binomial confidence intervals. The test is modified for the multiple pairwise comparison of four systems, and is applied to actual data to compare different configurations of the Moving and Stationary Target Acquisition and Recognition (MSTAR) System.

14. SUBJECT TERMS
Sequential Analysis, Multiple Comparisons, Probability, Binary Data Analysis, Methodology, Test and Evaluation, Automatic Target Recognition Systems

15. NUMBER OF PAGES
128

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL